Applications of Next Generation and Nanopore Sequencing for Infectious Disease Identification and Antimicrobial Resistance Detection

by
Stephanie Hao

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science and Engineering in the Department of Biomedical Engineering

Baltimore, Maryland

May 2017

Abstract:

Infectious diseases are a major problem worldwide, in no small part due to the growing prominence of antimicrobial resistance. Current methods for species identification rely on microbiology culturing techniques and nucleic acid tests, which are time consuming, may require some a priori knowledge of infectious agents, and are limited in the information provided. To address some of these limitations, clinical diagnostics laboratories have been applying shotgun DNA sequencing for disease detection. Traditionally, metagenomic sequencing directly from clinical specimens has not been as widely used in infectious disease, due to the costs associated with producing and analyzing the data. However, sequencing is now becoming more affordable and integrated into the clinical setting. One such example is the recently released MinION sequencer from Oxford Nanopore, a portable, low-cost sequencer that connects to standard personal computers via USB.

We are examining the application of the Oxford Nanopore MinION as a diagnostic aid for detecting pathogenic organisms in infectious disease, as well as acquisition of antimicrobial resistance. We hope to develop laboratory tests to identify and characterize infection causing organisms. The work for this master's thesis focused specifically on developing and optimizing sequencing and computational techniques that can be applicable to infectious disease diagnostics. We illustrate our progress using three separate cases: 1) detection of vancomycin and carbapenem resistance in pathogens from remnant rectal swabs, 2) a clinical case study involving an extensively drug resistant strain of *K. pneumoniae*, and 3) long-read sequencing of clinical influenza samples. We hope to leverage the MinION's versatility and sequence samples both from a clinical laboratory standpoint, as well as on site to locations of outbreaks.

Thesis Reviewers:

This thesis was reviewed and approved by the following faculty members (signatures are on file with the Department of Biomedical Engineering):

Winston Timp
Assistant Professor of Biomedical Engineering
Thesis Advisor and Chair

Patricia Simner
Assistant Professor of Pathology

Michael Schatz
Bloomberg Distinguished Associate Professor of Computer Science

Acknowledgements:

First off, I would like to thank my advisor, Dr. Winston Timp, for providing with me the opportunity to work in his lab, first as an undergraduate, and then as a Master's student in Biomedical Engineering. Over the past few years, I have been introduced to a variety of concepts, including Next Generation and Nanopore sequencing and bioinformatics analysis, especially in the context of disease diagnostics and microbiome research. Dr. Timp has also been a great mentor to me from both a personal and professional perspective.

I would also like to thank the members of the Timp Lab, especially to Rachael Sparklin, Isac Lee, Yunfan Fan, Jawara Allen, Tim Gilpatrick, and Evi Merken, for all of their guidance and support throughout my Master's work. I couldn't have asked for a better cohort to work with.

Next, I would like to thank all of our collaborators, first to Dr. Patricia Simner and Belita Opene, for teaching me about clinical diagnostics and microbial pathology, as well as for performing a lot of the clinical tests for many of the projects described in this thesis. I would also like to thank the folks at the Applied Physics Laboratory, especially to Peter Thielen and Tom Mehoke, for providing us with samples, guidance, and insight into bioinformatics analysis with the influenza project. In addition, I would like to thank the Salzberg lab, especially to Jennifer Lu and Florian Bretweiser, as well as to Dr.

Michael Schatz and James Gurtowski, for all of their advice regarding bioinformatics and computational analysis.

Finally, I would like to thank my thesis committee, for their guidance towards the projects I have worked on over the last two years, as well as their continual support.

Table of Contents

Table of Contents (Tables)

Table of Contents (Figures)

Chapter 1: Introduction

Infectious Disease and Antimicrobial Resistance Worldwide

Infectious disease mortality has decreased greatly over the last century; however, this trend has stagnated within the last 20-30 years (Dye, 2014). In 2010, 15 million people died of infectious diseases worldwide, and the World Health Organization estimates 13 million deaths per year in 2050 (Dye, 2014). This is in part due to the rise of pathogens resistant to traditional treatments such as antibiotics. Antimicrobial resistance is not a new concept; for instance, penicillin resistant Staphylococcus was discovered in the early 1940's (Lowy, 2003). However, it is a growing problem worldwide; for example, while penicillin-resistant *S. aureus* was only first identified in the 1940's, by the 1960's, 80% of all isolates of *S. aureus* were resistant (Lowy, 2003). Methilicilin was introduced to the general public around 1960, but only a few years later, methilicillin resistant strains of Staphylococcus were detected (Ventola, 2015). Vancomycin resistance first appeared in 1986, and by 1990's, vancomycin-resistant Enterococcus (VRE) were spreading throughout hospitals (O'Driscoll & Crank, 2015). Many bacteria are now multi-drug resistant (MDR) to a wide variety of antibiotics, and using only a single antibiotic for treatment is not enough. In a review article on carbapenemase-producing *Enterobacteriaceae*, up to 80% of all patients with carbapenem resistant organisms (CRO) treated with only tigecycline died, compared to only 50% mortality when treated with both tigecyline and gentamicin (Falagas, Lourida, Poulikakos, Rafailidis, & Tansarli, 2014). More recently, there have been cases of extremely drug resistant (XDR) pathogens, that have become more resistant to almost all antibiotics available in the clinic. In 2016, for example, there was a case of an extremely drug resistant *K. pneumoniae* infection resistant to every antibiotic approved by the Food and Drug

Adminstration in the United States, including colistin, one of the “last-resort” antibiotics (Chen, Todd, Kiehlbauch, Walters, & Kallen, 2017).

As of 2013, over 2 million people each year are infected with some form of drug resistant pathogen or virus In the United States, with at least 1 percent of those infections resulting in mortality (“Antibiotic Resistance Threats in the United States, 2013”, 2015). Patients infected with these pathogens often have longer hospital stays, and require more treatment and resources than those who are infected with susceptible strains (Ventola, 2015). Depending on the infection and treatment, these patients may have side/lasting effects that may require follow-up visits and treatment; in special circumstances, this may also involve a requirement for quarantine in order to not infect others (Flanagan et al., 2007). Unfortunately, pathogenic bacteria can easily be spread from person to person in a hospital setting, including to healthcare workers, visitors and between patients (Aitken & Jeffries, 2001). This is also a significant financial burden, as healthcare costs for hospitalization and treatment from AMR’s are estimated to cost $20 billion each year, not to mention the estimated $35 billion impact on the economy due to lost workforce (Smith & Coast, 2013).

The spread of antimicrobial resistance has become more prominent in recent years due to several major reasons. One major cause is the overuse of antibiotics, both in clinical and environmental settings. A recent study from the CDC found that nearly a third of all prescriptions were unnecessary, with many due to “overdiagnosis of certain conditions, such as sinusitis diagnosis without meeting criteria” (Fleming-Dutra et al., 2016). Additionally, patient compliance is also an issue, as they do not always take the full course of antibiotics that are necessary for treatment, stopping when they physically feel better or feel that they do not need them (Kardas, Devine, Golembesky, & Roberts, 2005) . In both well developed countries, including the United States, and developing countries, such as Brazil, Sudan, and Indonesia, non-

prescribed antibiotic use is high in the general population, where antibiotics could easily be acquired from either poorly regulated pharmacies or from family and friends (Morgan, Okeke, Laxminarayan, Perencevich, & Weisenberg, 2011). One of the unintended consequences of the excess use of antibiotics is the alteration of the human gut microbiome, with selections pressure for those species carrying antibiotics resistance genes. to maintain a healthy microbiome, promoting the growth of antibiotic resistant bacteria instead. Many of the genes responsible for antibiotic resistance are found on plasmids and/or flanked by transposable elements, permitting them to be transmitted from bacteria to bacteria, and thus further spreading the populations of resistant bacteria ("Antibiotic Resistance Threats in the United States, 2013", 2015).

Many antibiotics are also used for agriculture, as pesticides or to treat livestock to aid in growth and infection prevention. In fact, 80% of all antibiotic use in the United States is from agricultural use (Van Boeckel et al., 2015). These antibiotics can select for bacteria that are resistant to pathogens in certain environments, both directly, such as on the bacteria growing on the crops themselves, and indirectly, such as after animals ingest crops with the antibiotics. These bacteria can then spread relatively easily; for example, feces from animals carrying these antibiotic resistant bacteria can then get into the soil or water supply, which can be picked up by other organisms (Chang, Wang, Regev-Yochay, Lipsitch, & Hanage, 2015). Likewise, humans that are in close proximity to livestock can pick up the pathogens or viruses from the livestock or store bought meats. This has been seen in industrial hog workers, who have been seen to carry the same strains of multi-drug resistant *S. aureus* from nasal swabs that were found in the hogs that they worked with (Nadimpalli et al., 2015; Rinsky et al., 2013). These resistant bacteria can then be transferred to others, including family members (Rafee et al., 2012).

Unfortunately, the production of new antibiotics has stagnated. Thirty antibiotics were newly approved by the FDA in the 1980s; only 7 were approved in the 2000's

(Antibiotic/Antimicrobial Resistance, 2017). Fewer antibiotics are being studied and developed from a research perspective, both in academia and in industry. This is due to the perceived relative cost of research and return of investment - many older antibiotics are now available as a generic form, and are thus cheaply available. Newer antibiotics, on the other hand, tend now to be held off and used more as a "last resort" antibiotic, and thus not necessarily profitable to pharmaceuticals. The process of getting new antibiotics can be very time consuming and regulated, due to strong regulations and testing requirements from the FDA and other government entities (Power, 2006). The process for new drug approval typically takes about 8-12 years from laboratory tests all the way to being sold on the market and $802 million (DiMasi, Hansen, & Grabowski, 2003; Power, 2006). Only 1 out of 5000 drugs that start preclinical testing even make it to the human testing stage. This has led many large pharmaceutical companies to focus less on antibiotics; 80% of all antibiotic development is currently produced by small biotech companies, who do not have the resources that larger companies do ("Tracking the Pipeline of Antibiotics in Development," 2016).

Detection of Pathogenic Organisms in a Clinical Setting

Currently, clinical tests to detect infectious agents are primarily based on culture. This involves growing organisms in selective media, such as on an agar plate or in a liquid broth, then incubating at specific temperatures. Many organisms, such as *E. coli*, are rapid multipliers, and can be grown overnight at 37C in a lab. However, many organisms do not grow up under "standard" laboratory conditions; and some are unculturable. For example, urine was thought to be sterile due to lack of organism growth on a plate. More recently a group from Loyola University in Chicago discovered that greater than 90% of organisms found in urine are actually unculturable in a conventional microbiology lab (Nienhouse et al., 2014). A commonly employed

strategy is selectively growing for a particular trait, such as adding in an antibiotic to the culture media to test for resistance to that antibiotic, but this assumes prior knowledge of what to select for. When growing organisms in selective media, we can only select for specific characteristics at a time, such as a specific resistance to an antibiotic.

When looking for a specific gene, a targeted PCR approach may be employed. With this method, specific primers are designed that will amplify a region of interest. This method is much faster than culture; time to detection can occur within an hour, compared to the time necessary to culture up the sample. This method is also quite sensitive - one study could detect *Brucella* with concentrations as low as 10 colony forming units (CFU) using traditional PCR methods for selected targets, and even lower using RT-PCR (X. Wang et al., 2016). Another study using RT-PCR for influenza A viruses found 96% sensitivity for and 100% specificity (Richard K. Zimmerman, Charles R. Rinaldo, Mary Patricia Nowalk, G. K. Balasubramani, Mark G. Thompson, Arlene Bullotta, Michael Susick, Stephen Wisniewski, 2014) In addition, costs associated with such an assay are quite low per sample - a standard PCR can cost as low as $.20 per reaction, and qPCR can be as low as $1 per reaction (Purcell, Pearson, Frizelle, & Keenan, 2016). However, this method still has limitations. Genes mutate relatively quickly, especially in pathogenic bacteria and viruses. If the mutation occurs at the primer binding sequencing, false negatives are likely. Alternatively, false positives may be produced from mutations causing the primer to bind to other sites. PCR alone cannot determine if there are any variations within the gene that was amplified; making it desirable to have another analysis method for specific strain typing.

Other methods of disease detection in the clinical setting involve serological tests to probe host immune response, such as ELISA (enzyme-linked immunosorbent assays). ELISAs look for the presence of antibodies and/or antigens that are produced by the immune system as a response to a particular pathogen or virus. Antigens formed by diseases of interest are bound to

the surface of the plate, and are detected with fluorescently labelled antibodies. When these antibodies bind to the antigen, the fluorescence can be measured by a plate reader or some other fluorescent detector in order to determine the amount of antigen in the sample. ("Overview of ELISA," n.d.) This method is commonly used in high-throughput settings, as they are typically performed in 96 well plate/384 well reactions, and can be quite specific. Traditionally, this method is not quite as fast as targeted PCR, for example, due to various incubation steps required to prepare the antigen and antibodies throughout the process, and labelling multiple antibodies is time intensive. In addition, serological testing is not as sensitive in detection compared to PCR (Nilsson, Björkman, & Persson, 2008). There are point-of-care tests available based off of serological methods, but these tests can vary drastically in both sensitivity and specificity as well (Khuroo, Khuroo, & Khuroo, 2015).

Sequencing as used in Disease Detection

Instead of solely relying on culture based methods for diagnosis, clinicians have turned to sequencing techniques. DNA sequencing was first developed by Fred Sanger in the 1970's. Sanger sequencing works by appending a dNTP to the sequence using a polymerase to add on the next base pair. However, a small proportion of these dNTP's are replaced with ddNTP's, which when bound, will terminate sequence elongation since they lack the 3' hydroxyl group that binds to the 5' phosphate group in the next step (Heather & Chain, 2016). These samples are then run on a gel or capillary electrophoresis to determine the original sequence based on the terminating nucleotide. Today, Sanger sequencing is considered the gold standard of sequencing, and is widely accepted in the clinical setting, due to its relatively high accuracy, ability to call SNP's, and read lengths of up to 1000 base pairs. Even so, this method still has limitations. It requires a lot of starting material, which is problematic when clinical samples are

often limited. In addition, multiplexing samples is not practical with Sanger Sequencing due to difficulties with base calling; one would have to run separate reactions for each sample. Compared to other sequencing methods, Sanger sequencing is not as cheap per base pair (Chin, da Silva, & Hegde, 2013; Liu et al., 2012).

In the past decade, next generation sequencing (NGS) has gained prominence, transforming genetics and DNA analysis. Development through the National Human Genomic Research Institute along with private industry has brought the cost of sequencing the human genome from nearly three billion dollars to less than 1000 dollars. (Oetting, 2010; van Dijk, Auger, Jaszczyszyn, & Thermes, 2014). NGS differs from Sanger sequencing in 3 major ways: 1) It is no longer necessary to produce micrograms of sample by either plasmid cloning or through PCR, due to lower sample input requirements 2) More sequences can be run and more data generated and 3) Basecalling can be performed without the extra step of performing electrophoresis. However, the reads from NGS are shorter than with Sanger sequencing, which can make projects such as assembly difficult (van Dijk et al., 2014). Many biotech companies have developed NGS sequencers accessible to both laboratories and clinics. At the forefront is Illumina, whose sequencing by synthesis (SBS) technology leads the pack in terms of accuracy, throughput, and cost, and Thermo Fisher, whose Ion Torrent platform has also been found in the clinical setting due to its speed of sequencing and broad range of applications, including 16S sequencing for diagnosing bacterial infections (Salipante et al., 2013).

Common sequencing techniques employed in clinical settings include 16S sequencing, microarrays, and metagenomic shotgun sequencing. 16S sequencing is frequently used for pathogen identification. The 16S region is a highly conserved region of the bacterial genome comprising part of the prokaryotic ribosome and is unique between species, making pathogen identification relatively straightforward (Y. Wang, Tian, Gao, Bougouffa, & Qian, 2014).

Unfortunately, this region is only useful for species identification - information involving pathogen evolution, antibiotic resistance, or mutations is not provided which may impact treatment is unavailable. Additionally, 16S sequencing only works for bacteria; while other organisms, such as fungi, have conserved regions in their genome that can also be used for classification purposes, there is no consistent conserved region across different kingdoms.

Another common sequencing test performed in the clinical diagnostic world is the microarray. DNA samples from patients, as well as controls, are cut and labelled with fluorescent dyes. The sample is then added to a chip, which then hybridize to synthetic DNA oligos on the chip. If the sample binds to the microarray, then it will fluoresce - however, if there is a mutation, then it will not bind to the normal sample, but to another sequence that has the mutation, allowing us to detect SNP's in the sample ("DNA Microarray Technology," n.d.). This method is commercially available, and can be extremely effective for high-throughput screening, as multiple genes can be tested at the same time for relatively low cost. Doctors can order panels targeting a whole plethora of conditions, or choose a more targeted panel based on the symptoms that a patient may be presenting (Rehm, 2013). While this method has been applied mainly to cancer, it has also been utilized to diagnosing a variety of bacterial infections, such as multidrug resistant tuberculosis and S. aureus, and viruses, including HIV and influenza (Mikhailovich, Gryadunov, Kolchinsky, Makarov, & Zasedatelev, 2008).

However, there are still limitations to this method. The first is that mutations are constantly occurring in genes, and accounting for known mutations while anticipating new ones at the same time can be challenging. In addition, while there are genes with well-characterized mutation effects - such as specific mutations in the BRCA gene that lead to cancer, or extra CAG repeats in the HTT gene leads to Huntington's Disease - there are still many genes where even if a SNP is detected, it is unknown what cause that may actually have on acquiring a disease or

conferring resistance. These tests may also only determine the chance that one may develop a certain disease, as opposed to determining if you actually have the disease or not (Katsanis & Katsanis, 2013).

Unbiased next generation sequencing sidesteps many of the issues associated with culture, 16S sequencing and microarrays and unbiased NGS has become more cost-effective in the past decade. With this method, instead of sequencing only a specific region of a genome, everything in the sample is sequenced. *A priori* knowledge of what to look for is no longer required, and losing unculturable organisms is no longer a problem. However, there are still some limitations with this method. Coverage for particular infectious organisms may be lower than more targeted methods, which may be difficult when looking for mutations present in only a small amount of the input DNA. This can be fixed by performing deeper sequencing; despite the drop in price of sequencing over the years, this process can still be quite expensive, and often is limited to sequencing cores and centralized research centers. In addition, there is still variability between a lot of the different sequencers available, which provide varying levels of quality, length of read, genome coverage, and accuracy (Katsanis & Katsanis, 2013). One of the first applications of unbiased NGS for disease diagnosis was performed in 2014 to detect neuroleptospirosis in a 14 year old boy (Wilson et al., 2014), and is now being explored as a potential tool to be used in a clinical setting.

Nanopore Sequencing

Nanopore sequencing was first developed in the late 1980's by two different groups - one by David Deamer at the University of California, Santa Cruz, and the other by George Church from Harvard University (Deamer, Akeson, & Branton, 2016). Together, along with some of their colleagues, the two groups combined their ideas and patented it in 1995. Nanopore sequencing

involves sending a strand of nucleic acid through either a biological or synthetic pore on a membrane. A consistent voltage is passing through the membrane, and as the nucleic acid processes through the pore, the pore is blocked, changing the ion current. This change can be measured and used to determine what DNA nucleotides are passing through the pore at the time.

While many groups have been working on developing nanopore sequencers, one company, Oxford Nanopore, has become commercially successful. Founded in 2005 by Hagan Bayley, Spike Willcocks, David Norwood, and Gordon Sanghera, Oxford Nanopore first debuted their Nanopore MinION sequencer in 2012 at the Advances in Genome Biology and Technology conference. Their prototype was released to the early access group in 2014, and now is commercially available to all users (Deamer et al., 2016). DNA library preparation for the MinION involves the ligation of adapters containing a motor protein, responsible for ratcheting DNA through at a fixed rate, with 5-6 bases being read at a time (a k-mer). Because each k-mer has a unique current change, one can determine what the k-mer sequence was that went through the pore, and string together all the k-mers to determine the initial sequence (Figure 1).

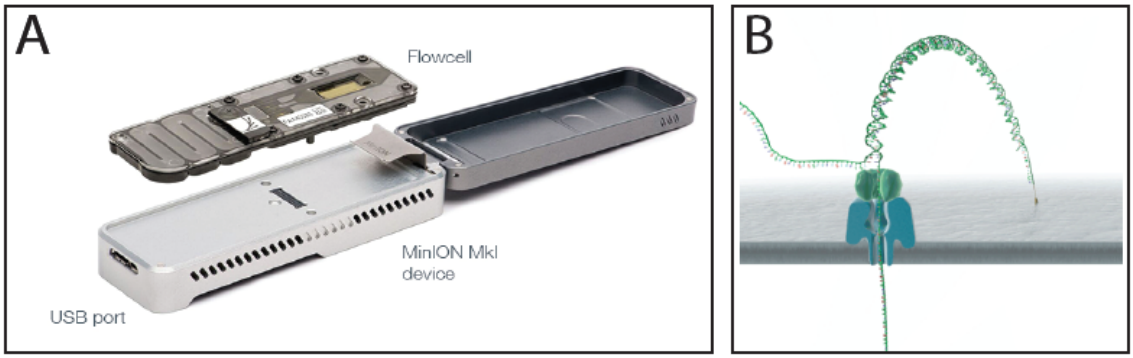

Figure 1: The Oxford MinION sequencer (A) Each strand of genomic material is passed through a pore (B). Sequences are determined based on the electrical current read out as the genomic material goes through the pore. MinION figure courtesy of Peter Thielen from the Johns Hopkins Applied Physics Laboratory and Oxford Nanopore Technologies for the pore.

Third generation sequencers, such as PacBio and Oxford Nanopore, have become increasingly more popular due to their ability to perform long-read sequencing. The current gold standard for sequencing, Illumina, is capable of producing sequencing reads between 25 and 600 basepairs in length. Oxford Nanopore, on the other hand, has been able to get reads as long as 771 kilobase pairs, with an average of 63800 base pair long reads ("Thar she blows! Ultra long read method for nanopore sequencing · Loman Labs," 2017). This can be extremely beneficial to projects that involve de novo assemblies, and determining the full genetic sequence of a gene region, for example.

Another advantage to the Nanopore sequencer is its portability. The MinION is a small sequencer that could fit into the palm of one's hand, and can be run on any computer connected via USB, including a laptop. This can be incredibly useful for sequencing at infectious disease outbreak sites, without a need for a complicated lab setup, and can be used in rural areas with limited resources. With new advances in basecalling methods, electrical signals can be translated to sequences even without an internet connection, making it even more feasible to take sequencing out onto the field or to extreme locations. Currently, the MinION has been tested both in controlled laboratory settings and out on the field in extreme environments, such as in forests, on lakes, and even on board the International Space Station.

Unlike other available sequencing technologies, the MinION can provide sequencing data in real-time, potentially allowing analysis to start once a sequence goes through the pore. This can be extremely beneficial in a healthcare setting, where treatment may rely on knowing what organisms may be causing the infection (for example, if the person is ill due to a viral versus a bacterial infection) and what else may be in your sample that can impact treatment. This can be done with Oxford's basecaller programs, such as What's In My Pot (WIMP), a k-mer based taxonomical classifier based off of Kraken that can show up to species level detection, or

can be analyzed by a separate outside analysis program, depending on what information you are looking for. In addition, newer technological and bioinformatic advances have also allowed for selective reading for particular samples, such as using read-until, which can sequence only the reads of pertinence and get rid of contaminants such as host reads (Loose, Malla, & Stout, 2016). By being able to have this real-time data, clinicians can provide more targeted, rapid treatment to patients.

Another benefit of nanopore sequencing is low capital cost. Currently, the MinION itself is only about $1000, a pittance compared to other sequencers, such as an Illumina Miseq ($125,000), PacBio Sequel ($350,000), and the Ion Torrent S5 ($65000) (Perkel & Fung, 2016). Consumables are also within the reach of many laboratories and are about on par with other companies; when bought in bulk, a single flow-cell can be as low as $500, and Oxford is working on improvements such that their sequencing consumables can become even cheaper. In addition, most of the library prep can be performed with basic lab equipment, such as pipettors and a heat block, which are commonly already found in most laboratory settings, and can be easily transported and used in more limited environments. This makes it relatively easier for laboratories to begin sequencing with lower overhead costs.

Nanopore sequencing has a potential for a wide range of sequencing applications, including disease detection and microbiome analysis. Work has already been performed on singular organisms/pathogens with relative success. One of the earliest experiments performed on the nanopore was to sequence reference bacterial genomes, FSC1006 and FSC996, on some of the earlier versions of the MinION flow cell (Karlsson, Lärkeryd, Sjödin, Forsman, & Stenberg, 2015). They found that overall error rates were relatively high, which is a problem that is common amongst third generation sequencers; however, due to higher coverage rates, consensus accuracy actually made it very comparable in variant calling with other sequencers

such as PacBio at 99.8% accuracy. Around the same time, another group performed de novo assembly of an *E. coli* genome into one full contig using Nanopore only data. Using just consensus, they got 98.4% accuracy compared to the reference genome, but after error correction using Nanopolish, an error corrector that utilizes the electrical data from the raw Nanopore reads, they got a more accurate consensus of 99.4% (Loman, Quick, & Simpson, 2015).

More recently, nanopore data has increased in yield and accuracy such that microbiome studies have become feasible. For example, a group in South Korea recently sequenced the mouse gut microbiome using full-length 16S rRNA sequencing using the MinION. Comparing their results to the Illumina Miseq, they found that the results were relatively correlated to each other, and that "nanopore sequencing was capable of determining the correct microbial composition up to the species level" (Shin et al., 2016). Another group from the University of Pittsburgh also evaluated the Nanopore sequencer for 16S rRNA microbiome characterization. While using reference based clustering is still problematic on the Nanopore due to high overall error rate, de novo clustering allowed for species calling to be comparable to an Illumina sequencer. (Ma, Stachler, & Bibby, 2017).

Some of the most notable disease detection work using the MinION out on the field was performed by the Loman group during the Ebola outbreak in 2014. Their team retrofitted a van into a mobile laboratory, including the capability of sequencing with the MinION. They then took this mobile lab and went to the locations of Ebola outbreaks, and began sequencing the cases that were occurring on site. They successfully tracked in real-time the different strains that were circulating throughout Africa and tracking the evolution of the virus over time (Quick et al., 2016). In 2015-2016, when the Zika virus was spreading, they once again set up a mobile lab

down in Brazil and began sequencing regions that were hit with the Zika virus to track its evolution (Faria et al., 2016).

Because of its ability to output results in real time, many groups have also been interested in using it in the clinical setting as a diagnostics method. A group at UCSF, for example, used the Nanopore MinION to sequence blood samples to identify Ebola, Hepatitis C, and Chikungunya Virus in less than 6 hours with their MetaPORE pipeline (Greninger et al., 2015). Zamin Iqbal's group used the MinION to perform whole genome sequencing of drug resistant tuberculosis, which could be taken from patients and sequenced to detect antimicrobial susceptibility within 7.5 hours, and full AST determined within 12 hours (Votintseva et al., 2017). In addition, a group from the Imperial College of London, spearheaded by Johanna Rhodes, has been using the Nanopore sequencer to examine the outbreak of *C. auris*, a pathogenic yeast that is also often multi-drug resistant ("Online Seminar: Jo Rhodes," 2017). Oxford Nanopore themselves are also working on examining pathogenic organisms streamlined into their pipeline; in addition to WIMP, they are also in the process of developing another workflow, ARMA, to determine antimicrobial resistance genes that are present in the sample.

Computational Resources to Detect Infectious Pathogens and Antimicrobial Resistance

The most widely used computational method for detecting for a specific pathogen or AMR gene from sequencing data is to compare it to a reference genome or database. This is commonly done through programs such as the Basic Local Alignment Search Tool (BLAST). BLAST

works by breaking down sequences into “words”, or segments of specific lengths determined by the user, and comparing those to similar length words in the database. After BLAST finds a similar match, the program extends the alignment in either direction until the match score falls below a certain threshold (Altschul, Gish, Miller, Myers, & Lipman, 1990; “Basic Local Alignment Search Tool (BLAST) | Learn Science at Scitable,” n.d.).

For metagenomic sequencing, one of the most commonly used programs for species identification is to use Kraken, a k-mer based taxonomic classifier. Kraken works by taking a query sequence and extracting every k-length subsequence from the query. Each k-mer is then compared to the database and assigned to the lowest-common ancestor of all genomes with an exact match for that k-mer. The query sequence is then assigned a classification based on all k-mer assignments. (Wood & Salzberg, 2014). Other similar programs include MEGAN, which first blasts all of the sequences against a database and reports back the lowest common ancestor (Huson, Auch, Qi, & Schuster, 2007), and MetaPhIAn, a taxonomic classifier that compares sequences to a “reduced set of clade-specific marker sequences” to identify the microbial community (Segata et al., 2012).

However, all of these programs rely on the use of databases. Many of these programs rely on BLAST and therefore use databases from the National Center for Biotechnology Information (NCBI). This comprehensive collection of databases includes nucleotide and protein sequences for all types of organisms, including bacteria, fungi, and eukaryotes. Databases can also be customized based on the application that is at hand. In the case of antimicrobial resistance, for example, there are multiple databases available, such as the Comprehensive Antibiotic Resistance Database (CARD) (Jia et al., 2017), Resfinder (Zankari et al., 2012), and MEGARes (Lakin et al., 2017). One of the major challenges, however, is to keep the databases

updated frequently in order to account for the influx of sequencing data and new species and resistance genes that are being discovered.

In this thesis, we describe our work on applying nanopore sequencing in the clinical setting. The following chapters describe the work performed in validating the MinION for detecting infectious pathogens and antimicrobial resistance, from both the sequencing and the bioinformatics perspective.

References:

Aitken, C., & Jeffries, D. J. (2001). Nosocomial spread of viral disease. *Clinical Microbiology Reviews*, *14*(3), 528–546.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Antibiotic / Antimicrobial Resistance. (2017, March 08). Retrieved April 10, 2017, from https://www.cdc.gov/drugresistance/cdc_role.html

"Antibiotic Resistance Threats in The United States, 2013| Antibiotic/Antimicrobial Resistance | CDC". (2015) Retrieved March 10, 2017, from https://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf

Basic Local Alignment Search Tool (BLAST) | Learn Science at Scitable. (n.d.). Retrieved May 2, 2017, from https://www.nature.com/scitable/topicpage/basic-local-alignment-search-tool-blast-29096

Chang, Q., Wang, W., Regev-Yochay, G., Lipsitch, M., & Hanage, W. P. (2015). Antibiotics in agriculture and the risk to human health: how worried should we be? *Evolutionary Applications*, *8*(3), 240–247.

Chen, L., Todd, R., Kiehlbauch, J., Walters, M., & Kallen, A. (2017). Notes from the Field: Pan-Resistant New Delhi Metallo-Beta-Lactamase-Producing Klebsiella pneumoniae - Washoe County, Nevada, 2016. *MMWR. Morbidity and Mortality Weekly Report*, *66*(1), 33.

Chin, E. L. H., da Silva, C., & Hegde, M. (2013). Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. *BMC Genetics*, *14*, 6.

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5), 518–524.

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, *22*(2), 151–185.

DNA Microarray Technology. (n.d.). Retrieved May 3, 2017, from https://www.genome.gov/10000533/dna-microarray-technology/

Dye, C. (2014). After 2015: infectious diseases in a new era of health and development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1645), 20130426.

Falagas, M. E., Lourida, P., Poulikakos, P., Rafailidis, P. I., & Tansarli, G. S. (2014). Antibiotic treatment of infections due to carbapenem-resistant Enterobacteriaceae: systematic evaluation of the available evidence. *Antimicrobial Agents and Chemotherapy*, *58*(2), 654–663.

Faria, N. R., Sabino, E. C., Nunes, M. R. T., Alcantara, L. C. J., Loman, N. J., & Pybus, O. G. (2016). Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, *8*(1), 97.

Flanagan, M., Ramanujam, R., Sutherland, J., Vaughn, T., Diekema, D., & Doebbeling, B. N. (2007). Development and validation of measures to assess prevention and control of AMR in hospitals. *Medical Care*, *45*(6), 537–544.

Fleming-Dutra, K. E., Hersh, A. L., Shapiro, D. J., Bartoces, M., Enns, E. A., File, T. M., Jr, … Hicks, L. A. (2016). Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. *JAMA: The Journal of the American Medical Association*, *315*(17), 1864–1873.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., … Chiu, C. Y. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Medicine*, *7*, 99.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., … McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *45*(D1), D566–D573.

Kardas, P., Devine, S., Golembesky, A., & Roberts, C. (2005). A systematic review and meta-analysis of misuse of antibiotic therapies in the community. *International Journal of Antimicrobial Agents*, *26*(2), 106–113.

Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., & Stenberg, P. (2015). Scaffolding of a bacterial genome using MinION nanopore sequencing. *Scientific Reports*, *5*, 11996.

Katsanis, S. H., & Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nature Reviews. Genetics*, *14*(6), 415–426.

Khuroo, M. S., Khuroo, N. S., & Khuroo, M. S. (2015). Diagnostic accuracy of point-of-care tests for hepatitis C virus infection: a systematic review and meta-analysis. *PloS One*, *10*(3), e0121450.

Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., … Boucher, C. (2017). MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Research*, *45*(D1), D574–D580.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., … Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, *2012*, 251364.

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, *12*(8), 733–735.

Loose, M., Malla, S., & Stout, M. (2016). Real-time selective sequencing using nanopore technology. *Nature Methods*, *13*(9), 751–754.

Lowy, F. D. (2003). Antimicrobial resistance: the example of Staphylococcus aureus. *The Journal of Clinical Investigation*, *111*(9), 1265–1273.

Ma, X., Stachler, E., & Bibby, K. (2017, January 29). *Evaluation of Oxford Nanopore MinION Sequencing for 16S rRNA Microbiome Characterization*. *bioRxiv*. https://doi.org/10.1101/099960

Mikhailovich, V., Gryadunov, D., Kolchinsky, A., Makarov, A. A., & Zasedatelev, A. (2008). DNA microarrays in the clinic: infectious diseases. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *30*(7), 673–682.

Morgan, D. J., Okeke, I. N., Laxminarayan, R., Perencevich, E. N., & Weisenberg, S. (2011). Non-prescription antimicrobial use worldwide: a systematic review. *The Lancet Infectious Diseases*, *11*(9), 692–701.

Nadimpalli, M., Rinsky, J. L., Wing, S., Hall, D., Stewart, J., Larsen, J., … Heaney, C. D. (2015). Persistence of livestock-associated antibiotic-resistant Staphylococcus aureus among industrial hog operation workers in North Carolina over 14 days. *Occupational and Environmental Medicine*, *72*(2), 90–99.

Nienhouse, V., Gao, X., Dong, Q., Nelson, D. E., Toh, E., McKinley, K., … Radek, K. A. (2014). Interplay between bladder microbiota and urinary antimicrobial peptides: mechanisms for human urinary tract infection risk and symptom severity. *PloS One*, *9*(12), e114185.

Nilsson, A. C., Björkman, P., & Persson, K. (2008). Polymerase chain reaction is superior to serology for the diagnosis of acute Mycoplasma pneumoniae infection and reveals a high rate of persistent infection. *BMC Microbiology*, *8*, 93.

O'Driscoll, T., & Crank, C. W. (2015). Vancomycin-resistant enterococcal infections: epidemiology, clinical manifestations, and optimal management. *Infection and Drug Resistance*, *8*, 217–230.

Oetting, W. S. (2010). Impact of next generation sequencing: the 2009 Human Genome Variation Society Scientific Meeting. *Human Mutation*, *31*(4), 500–503.

Online Seminar: Jo Rhodes. (n.d.). Retrieved April 22, 2017, from https://register.nanoporetech.com/online-seminar-jo-rhodes

Overview of ELISA. (n.d.). Retrieved from https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-elisa.html

Perkel, J. M., & Fung, P. A. (2016, February 18). Next-Gen Sequencing 2016 Update. Retrieved May 3, 2017, from http://www.biocompare.com/Editorial-Articles/183239-Next-Gen-Sequencing-2016-Update/

Power, E. (2006). Impact of antibiotic restrictions: the pharmaceutical perspective. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *12 Suppl 5*, 25–34.

Purcell, R. V., Pearson, J., Frizelle, F. A., & Keenan, J. I. (2016). Comparison of standard, quantitative and digital PCR in the detection of enterotoxigenic Bacteroides fragilis. *Scientific Reports*, *6*, 34554.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232.

Rafee, Y., Abdel-Haq, N., Asmar, B., Salimnia, T., Pharm, C. V., Rybak Pharm, M. J., & Amjad, M. (2012). Increased prevalence of methicillin-resistant Staphylococcus aureus nasal colonization in household contacts of children with community acquired disease. *BMC Infectious Diseases*, *12*, 45.

Rehm, H. L. (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nature Reviews. Genetics*, *14*(4), 295–300.

Richard K. Zimmerman, Charles R. Rinaldo, Mary Patricia Nowalk, G. K. Balasubramani, Mark G. Thompson, Arlene Bullotta, Michael Susick, Stephen Wisniewski. (2014). Detection of Influenza Virus Infection Using Two PCR Methods. *Advances in Virology*, *2014*. https://doi.org/10.1155/2014/274679

Rinsky, J. L., Nadimpalli, M., Wing, S., Hall, D., Baron, D., Price, L. B., … Heaney, C. D. (2013). Livestock-associated methicillin and multidrug resistant Staphylococcus aureus is present among industrial, not antibiotic-free livestock operation workers in North Carolina. *PloS One*, *8*(7), e67641.

Salipante, S. J., Sengupta, D. J., Rosenthal, C., Costa, G., Spangler, J., Sims, E. H., … Hoffman, N. G. (2013). Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PloS One*, *8*(5), e65226.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811–814.

Shin, J., Lee, S., Go, M.-J., Lee, S. Y., Kim, S. C., Lee, C.-H., & Cho, B.-K. (2016). Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing. *Scientific Reports*, *6*, 29681.

Smith, R., & Coast, J. (2013). The true cost of antimicrobial resistance. *BMJ* , *346*, f1493.
Thar she blows! Ultra long read method for nanopore sequencing · Loman Labs. (n.d.). Retrieved April 22, 2017, from http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/?rev1
Tracking the Pipeline of Antibiotics in Development. (2016, December 21). Retrieved May 3, 2017, from http://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2014/03/12/tracking-the-pipeline-of-antibiotics-in-development
Van Boeckel, T. P., Brower, C., Gilbert, M., Grenfell, B. T., Levin, S. A., Robinson, T. P., … Laxminarayan, R. (2015). Global trends in antimicrobial use in food animals. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(18), 5649–5654.
van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics: TIG*, *30*(9), 418–426.
Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *P & T: A Peer-Reviewed Journal for Formulary Management*, *40*(4), 277–283.
Votintseva, A. A., Bradley, P., Pankhurst, L., Del Ojo Elias, C., Loose, M., Nilgiriwala, K., … Iqbal, Z. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples. *Journal of Clinical Microbiology*. https://doi.org/10.1128/JCM.02483-16
Wang, X., Li, X., Liu, S., Ren, H., Yang, M., Ke, Y., … Chen, Z. (2016). Ultrasensitive Detection of Bacteria by Targeting Abundant Transcripts. *Scientific Reports*, *6*, 20393.
Wang, Y., Tian, R. M., Gao, Z. M., Bougouffa, S., & Qian, P.-Y. (2014). Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PloS One*, *9*(3), e90053.
Wilson, M. R., Naccache, S. N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., … Chiu, C. Y. (2014). Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *The New England Journal of Medicine*, *370*(25), 2408–2417.
Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46.
Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., … Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of Antimicrobial Chemotherapy*, *67*(11), 2640–2644.

Chapter 2: Vancomycin Resistance and Carbapenem Resistant Organism Detection from Perirectal Swab Samples

Authors Note and Acknowledgements:

The following manuscript describes our work in performing metagenomic shotgun sequencing from peri-rectal swab samples to detect vancomycin resistant and carbapenem resistant organisms using both the Illumina and Nanopore sequencing platforms. For this project, I performed all of the genomic DNA extraction, sequencing, and the majority of the sequencing data analysis for this project; however, this would not have been possible without Belita Opene, who performed the clinical aspects for this study, Rachael Workman, who helped teach me both Illumina and Nanopore sequencing, and of course, to Dr. Simner and Dr. Timp, who had initially conceived the study, and helped guide me through the project. I would also like to thank Florian Bretweiser for helping us develop the customized Kraken database for antibiotic resistance detection.

This is currently a manuscript that is in progress, and is expected to be published sometime in 2017/2018. The citation is as follows:

Hao, S., Opene, B., Workman, R., Simner, P., Timp, W. (2017). *Shotgun Metagenomic Sequencing of Perirectal Swabs for Surveillance of Antimicrobial Resistant Organisms on the Illumina Miseq and Oxford Nanopore MinION.* Manuscript in Preparation. Johns Hopkins University

Shotgun Metagenomic Sequencing of Perirectal Swabs for Surveillance of Antimicrobial Resistant Organisms on the Illumina Miseq and Oxford MinION

Stephanie Hao[1], Belita Opene[2], Rachael Workman[1], Patricia Simner[2] and Winston Timp[1]

Affiliations:

1 Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland

2 Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD

Abstract

Antimicrobial resistance (AMR) is an emerging epidemic problem worldwide. Existing methods for detection of gastrointestinal colonization are either time-consuming or too targeted to allow for rapid response and clear understanding of transmission mechanism. Here we outline our efforts to apply shotgun metagenomic sequencing, using both the Illumina Miseq and Oxford Nanopore MinION, to characterize vancomycin and carbapenem resistance from patient perirectal swabs.

Introduction

Antimicrobial resistant organisms (ARO) that are known to colonize the gastrointestinal tract, such as vancomycin-resistant enterococci (VRE) and carbapenem-resistant Gram-negative organisms (CRO), are associated with significant morbidity and mortality especially among immunocompromised populations (Arias, Contreras, and Murray 2010), (Nordmann, Naas, and Poirel 2011). Early detection of ARO colonization is highly beneficial in the healthcare setting as patients with unrecognized colonization and/or infection with ARO serve as a reservoir for transmission during health-care associated outbreaks (Snitkin et al. 2012; Samra et al. 2007). As such, certain hospital institutions have opted to implement ARO surveillance programs in high-

risk populations such as those in intensive care units (ICUs), oncology and transplant wards. These surveillance programs aim to identify patients with gastrointestinal colonization of AROs to place these patients on contact precautions to prevent the spread in the hospital setting to other vulnerable patients. In addition, therapeutic options for infections caused by these AROs are limited as they are often multidrug-resistant (MDR). This direct clinical application highlights how early and accurate identification of AROs is important given the implications for infection control measures and early and appropriate antibiotic therapy.

Current laboratory methods for the detection of gastrointestinal colonization with VRE and CRO are based on selective culture based techniques or PCR based assays from perirectal swabs (Simner et al. 2016; Suwantarat et al. 2014; Amberpet et al. 2016). Both methods are narrowly targeted towards a specific ARO type and PCR based methods may be limited further by the targets included in the assay. With the growing number of antimicrobial resistance genes, such as the recent emergence of plasmid-mediated colistin resistance (Liu et al. 2016), it will be difficult for Clinical Microbiology Laboratories to continue to perform targeted based methods for each ARO type encountered. Thus, a broader based approach to detect all AROs is required.

One such approach is the use of shotgun metagenomic analysis of specimens using next-generation sequencing platforms. Shotgun analysis of antimicrobial resistance genes in clinical specimens by Next-Generation Sequencing (NGS) is an application that can overcome current limitations in resistance gene detection. Whole genome shotgun sequencing has the advantage over targeted NGS by amplifying any DNA present in the sample versus the use of targeted primers that amplifies specific organism groups (i.e., 16S rDNA for bacteria) or resistance genes. This allows for the ability to not only query the entire microbiome of the sample (bacteria, DNA viruses, fungi and parasites), but also the presence of resistance genes to detect and study the

resistome. Theoretically, we can also obtain genotypic information about the exact variant of AMR genes; single nucleotide polymorphism (SNP) analysis may even allow us to trace infections of the same agent between patients. These advances have clinical implications with the ability to allow for tailored antimicrobial use and inform rapid implementation of infection control measures for a broad range of AROs.

Different types of sequencing technology offer distinct advantages for infectious agent surveillance and diagnosis. Illumina's sequencing-by-synthesis (SBS) technology is the current market leader, with high accuracy and throughput (Loman et al. 2012). A single run can allow millions of individual sequencing reads, giving high confidence on detection of different organisms. More recently, nanopore sequencing has been commercialized in the form of the Oxford Nanopore MinION. Nanopore sequencing directly reads DNA by detecting the the ionic current as DNA that passes through a nanoscopic pore (Timp et al. 2014). Though the ability to call single base mutations is not as accurate as Illumina sequencing, the reads are much longer - allowing for identifying the location and context of the genes (e.g. plasmid or main genome). Long reads also allow for analysis through repetitive DNA sequences associated with AMR genes and mobile genetic elements; this is difficult with short reads due to poor mappability in these regions (Ashton et al. 2015). Further, Nanopore sequencing also provides results in real-time, allowing for classification on the time scale of minutes, with more evidence accumulating throughout the sequencing run (Cao et al. 2016). There is low capital expenditure and low infrastructure overhead, enabling the potential distribution of this technology for widespread clinical use.

The purpose of this study was threefold: 1) to determine if shotgun metagenomic NGS analysis can be applied to rectal swabs, 2) to determine if the results are comparable to culture based methods for the surveillance of ARO such as VRE and CRO and 3) to compare the performance

of both the Illumina MiSeq and the MinION technologies. As such, we applied both of these sequencing technologies to a series of remnant clinical VRE surveillance rectal swabs to demonstrate the depth of information available. These results were compared to the results from the standard-of-care clinical microbiology testing. We also performed control experiments with spike-ins of our AROs of interest (VRE and CRO) into a negative control sample, to measure the semi-quantitative ability of this technology.

Our results show that sequencing allows for the rapid measurement of the full microbiome and resistome of samples, correlating well with each other and the AST testing while providing much greater depth of information. We were able to characterize the exact variant of AMR genes present in the sample, including both the most prevalent form of the AMR gene as well as others not found through traditional culture based methods. Finally, using long reads from Nanopore sequencing, we could identify certain AMR genes within minutes of starting the run, and further, we identified the context of AMR genes and demonstrated rapid microbiome and resistome mapping.

Results:

Standard Care ARO Surveillance Culture Results:

To provide a general assessment of shotgun metagenomic NGS applications on clinical specimens, we obtained ten remnant peri-rectal swabs that were evaluated for vancomycin-resistant enterococci (VRE) and carbapenem-resistant organism (CRO) surveillance cultures. VRE surveillance cultures were performed using the VRE Select™ (BioRad, Hercules, CA) chromogenic agar and CRO surveillance cultures were performed by a direct MacConkey plate method with ertapenem disks as previously described (Simner et al. 2016). Through

this method we semi-quantified and described the growth of Gram-negative organisms observed (1-4+ growth; lactose fermenter versus non-lactose fermenter) and the presence of potential CROs based off the zone of inhibition of the organisms around the ertapenem disks. Any isolate that grew within 27 mm of the ertapenem disks were further identified by matrix-assisted laser-desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS; Bruker Daltonics Inc., Billerica, MA) and antimicrobial susceptibility testing. (AST) was performed by the BD Phoenix™ Automated Instrument (Becton Dickinson, Sparks, MD, USA) and interpreted following Clinical and Laboratory Standards Institute's Guidelines. *Enterobacteriaceae* resistant to one of the carbapenems tested (ertapenem and meropenem) or glucose non-fermenting Gram-negative bacilli resistant to meropenem were further evaluated by the Carba NP assay to identify carbapenemase production. If an isolate was determined to be positive by the Carba NP, the molecular genotype was determined by the Check-MDR CT103XL assay as previously described (Check-Points, Wageningen, the Netherlands). Culture based results are summarized in Table 1. Most studies on the application of metagenomic NGS as a diagnostic tool from clinical specimens have focused on sterile site specimens such as CSF, tissue and blood (Wilson et al. 2014; Naccache et al. 2015; Berg et al. 2015). Instead, we wanted to test using a complex specimen source (i.e. peri-rectal swabs) where data analysis is complicated by inhibitors and large numbers of diverse normal gastrointestinal microorganisms.

Table 1: Overall results for all samples for culture and sequencing. VRE and CRO results by culture, as well as metagenomic shotgun sequencing results from both the Illumina Miseq and the Oxford Nanopore MinION

	Vancomycin-resistant Enterococci Culture Results		Carbapenem-Resistant Organism Culture Results		Miseq Data		Nanopore Data	
VRE#	VRE chromogenic culture results	Organism based on chromogen (pink colonies = E. faecium; blue colonies = E. faecalis)	Growth on MacConkey Agar with ertapenem disk	Carbapenem resistant organism (CRO) culture results	Dominant bacterial organisms	Resistance	Dominant Bacterial Organisms	Resistance
NC	Negative for VRE	Negative	No Growth	Negative	*K. oxytoca, E. faecalis, E. coli*	no detection	*K. oxytoca, E. faecalis*	no detection
1	Positive for VRE	*Enterococcus feacium*	4 + Lactose fermenter	Negative	*K. pneumoniae, P. difficile, E. faecium*	*vanA* positive	not run	not run
2	Positive for VRE	*Enterococcus faecalis*	4+ Mixed lactose fermenters	Negative	*E. cloacae, P. distasonis, E. asburiae*	*vanA* (<15 hits) and *vanB* positive	not run	not run
3	Positive for VRE	*Enterococcus feacium*	4+ Non-lactose fermenter; no zone around ertapenem disk	Positive: Meropenem resistant Pseudomonas aeruginosa	*P. aeruginosa, E. faecalis, E. faecium*	*vanA* positive	*P. aeruginosa*	*vanA* positive
4	Positive for VRE	*Enterococcus feacium*	4+ Non-lactose fermenter; 21 mm zone around ertapenem disk	Positive: Ertapenem resistant Enterobacter cloacae	*E. cloacae, E. asburiae, C. sakazakii*	*vanA* positive (<15 hits)	not run	not run
5	Positive for VRE	*Enterococcus feacium*	No Growth	Negative	*E. faecium, S. epidermitis*	*vanA* positive	*E. faecium, S. epidermitis*	*vanA* positive
6	Positive for VRE	*Enterococcus feacium*	1+ Non-lactose fermenter; >40 mm zone around ertapenem disk	Negative	*Achromobacter xylosoxidans, E. faecium*	*vanA* positive	*Achromobacter xylosoxidans, E. faecium*	*vanA* positive
7	Positive for VRE	*Enterococcus feacium*	1+ Lactose fermenter; 12 mm zone around ertapenem disk	Positive: KPC-producing K. pneumoniae; Check-Direct CPE multiplex PCR assay: Ct 14.1 for KPC	*K. pneumoniae*	*vanA , vanB , KPC* positive	*K. pneumoniae*	*vanA , KPC* positive
8	Positive for VRE	*Enterococcus feacium*	Mixed Lactose fermenters; 36 mm zone around ertapenem disk	Negative	*K. pneumoniae, E. faecalis*	*vanA* positive	not run	not run
9	Positive for VRE	*Enterococcus feacium*	No Growth	Negative	*E. faecium*	*vanA* positive	not run	not run
10	Positive for VRE	*Enterococcus faecalis*	Non-lactose fermernter ("Pseudo"); 39 mm zone around ertapenem	Negative	*Parabacteroides distasonis, Enterobacter*	*vanA , vanB, KPC* positive	*Parabacteroides distasonis, Enterobacter*	*vanB* positive

Illumina Sequencing

Microbiome Characterization

We isolated gDNA from 10 remnant peri-rectal swab samples and a negative control sample from pooled rectal swabs. Nextera libraries for each sample were created and sequenced on the Illumina MiSeq, multiplexing ~3 samples per run. We produced an average of 9.7 Million paired-end reads per sample (6.0M-15.1M range), for an average of 1.4 Gb sequence (0.9-2.2Gb range). To characterize the microbiome from these reads, we used Kraken - a rapid taxonomic classification algorithm (Wood and Salzberg 2014). Table 1 shows an overview of our samples comparing VRE and CRO surveillance culture results and sequencing data.

Our Illumina sequencing results were concordant with the semi-quantitative culture results. As detailed in Table 2 and Figure 2, we characterized the microbiome of these samples using Kraken, showing human, unclassified, viral/fungal, and dominant bacterial species. Overall, human reads were relatively low on the VRE positive swab samples, ranging between 0.14% and 8.33% for 9 of the 10 samples with the lone exception of VRE8 (41%). Our negative control had more than 50% of all reads identified as human; possibly due to an artifact from the pooled samples used there. This wide range of values are on par with other studies such as Vincent et al, who found that human values when looking at fecal samples in hospitalized patients, ranged from 0.1-98% of total reads (Vincent et al. 2016).

Table 2: Percent of Bacterial and host reads identified through sequencing. For each of the samples that were sequenced on the Illumina Miseq, we report the % of reads that were bacterial and the % of reads that were identified as human. In addition, we report the percentage of bacterial reads that were detected to be either E. faecalis, E. faecium, K. pneumoniae, and Other Bacteria.

Control	% Bacterial Reads	% Human Reads	% Bacteria ENFS	% Bacteria ENFM	% Bacteria KPN	% Other Bacteria
ENFM	87.08%	0.00%	0.16%	81.57	0.09	5.278184238
ENFS	98.31%	0.00%	97.72%	0.07	0.02	2.14049725
KPN	96.34%	0.00%	0.03%	0.01	85.99	10.30146852
NC	21.04%	70.71%	12.60%	0.06	0.25	18.08295454
PC1	21.75%	50.81%	12.56%	0.05	2.96	16.00692077
PC2	25.55%	50.72%	28.57%	0.1	0.24	17.91063377
PC3	26.09%	45.00%	14.05%	3.32	0.32	18.78292531
VRE 1	60.06%	8.33%	3.34%	0.06	45.52	12.46858356
VRE 2	55.04%	0.02%	0.11%	0.05	0.61	54.31547446
VRE 3	89.40%	5.63%	2.19%	0.89	0.03	86.52701526
VRE 4	50.59%	7.13%	0.00%	0.02	1.03	49.53870866
VRE 5	47.17%	0.53%	0.05%	23.6	0	23.54242965
VRE 6	33.83%	0.75%	0.07%	13.5	0.01	20.29367937
VRE 7	44.71%	0.50%	0.00%	0.23	38.36	6.116609146
VRE 8	46.82%	41.19%	42.43%	0.7	21.18	5.083143662
VRE 9	33.07%	2.52%	22.18%	23.51	0.01	2.21934731
VRE 10	48.92%	0.14%	0.36%	0.07	0.41	48.26413646

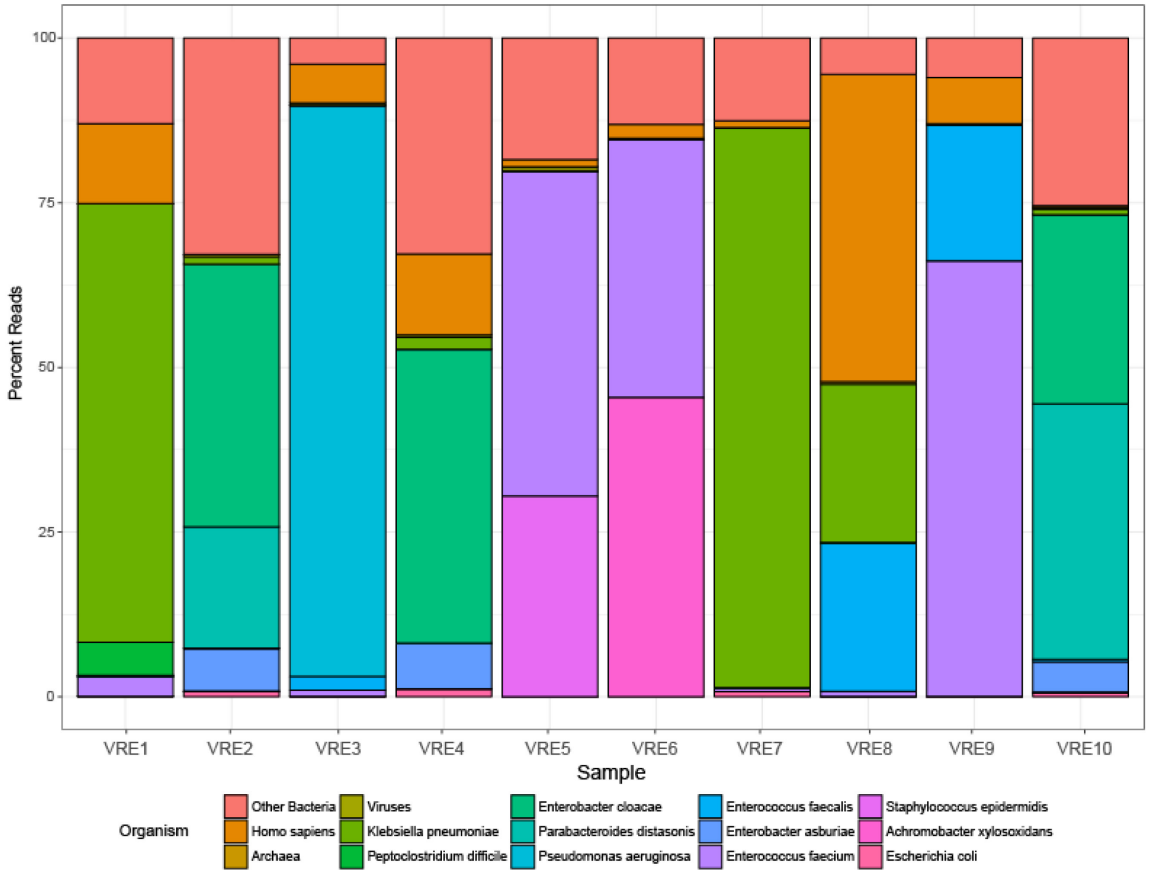

Figure 2: Metagenomic characterization of each clinical sample. For each sample, we examined the Bacterial, Viral, and host composition, and plotted their distribution in a stacked bar chart. Note that unidentified organisms are not plotted.

By culture, all 10 rectal swabs were confirmed to be positive for VRE, with 2 samples positive for *Enterococcus faecalis* (VRE2, VRE10), and the rest positive for *Enterococcus faecium*. In addition, 3 of the samples (VRE3, VRE4, and VRE7) were positive for carbapenem-resistant organisms, with VRE3 being positive for *Pseudomonas aeruginosa*, VRE4 positive for *Enterobacter cloacae*, and VRE7 positive for *Klebsiella pneumoniae*. Our negative control sample was neither positive for VRE nor CRO. The top dominant organisms for each sample are listed in Table 1.

We were able to detect the same organisms found by chromagen results, as well as the corresponding antibiotic resistance gene found for *E. faecium* (ENFM) and *E. faecalis* (ENFS), with the exception of VRE4, where no vancomycin resistance genes were detected. This is not

surprising, as chromagen results also showed lower growth of VRE than the other samples. VRE7 was also found to encode the gene ***Klebsiella pneumoniae carbapenemase*** (*KPC*) by growing on MacConkey Agar with ertapenem disks, and sequencing was also able to detect high quantities of the KPC gene, with over 7000 hits.

No growth occurred on MacConkey agar for samples VRE5 and VRE9, because the primary organism is a Gram-positive bacteria (*E. faecium*). This exemplifies the problem with culture based methods for surveillance and diagnosis; as sequencing does not depend on any of these selection methods to obtain results, we can easily profile the microbiome without worrying about the culturability of the organisms. It is also often assumed that no growth on MacConkey agar from rectal swabs is due to inappropriate collection but may rather reflect a predominance of Gram-positive organisms among the gastrointestinal microbiome. It would be interesting to link these sample back to antibiotic use. One can hypothesize that these patients were exposed to Gram-negative agents selecting for Gram-positive organisms among the gastrointestinal microbiome, such as *Enterococcus* spp.

K. pneumoniae was extremely abundant in VRE1 and VRE7, making up nearly 40% of all reads, and relatively abundant in VRE8, which took up about 20% of all reads. ENFS and ENFM are not quite as abundant in our samples, with only three samples having greater than 10% ENFM (samples VRE5, VRE6, and VRE9). ENFS only had two samples with a notable amount of ENFS, VRE8 and VRE9. With the majority of samples, VRE related organisms typically comprised of less than 1% of all reads; other organisms dominated the microbiome, unsurprisingly, given the gut microbiome diversity in humans (Eckburg et al. 2005), (Budding et al. 2014).

Organisms, antibiotic resistance detection

We used a customized Kraken database built from the CARD database (McArthur et al. 2013) to detect antibiotic resistance. An example heatmap of our results is shown below in Figure 3. Using this customized Kraken database, we were able to detect *vanA* in all of our samples, although some of our samples had relatively low quantities. For example, with sample 2, we only had 12 hits for *vanA*, and with VRE4, only 1 hit was *vanA*. With such low values, these may be false positives. VRE2, for example, was positive for ENFS by clinical testing, which should have possessed the *vanB* gene. We were also able to detect *vanB* in three of our samples, 2, 7, and 10, with varying levels of detection from 24-179 hits. KPC was determined in two of our samples, 7 and 10. Sample 7 had over 8500 reads that were determined to be *KPC*, which is consistent with clinical findings for CRO. Sample 10, on the other hand, only had 3 hits for the *KPC* gene, which may be carryover due to multiplexing both samples 7 and 10 on the same Illumina cartridge. While there were many hits for *KPC* in VRE7, we were unable to identify specific genotype of the *KPC* gene.

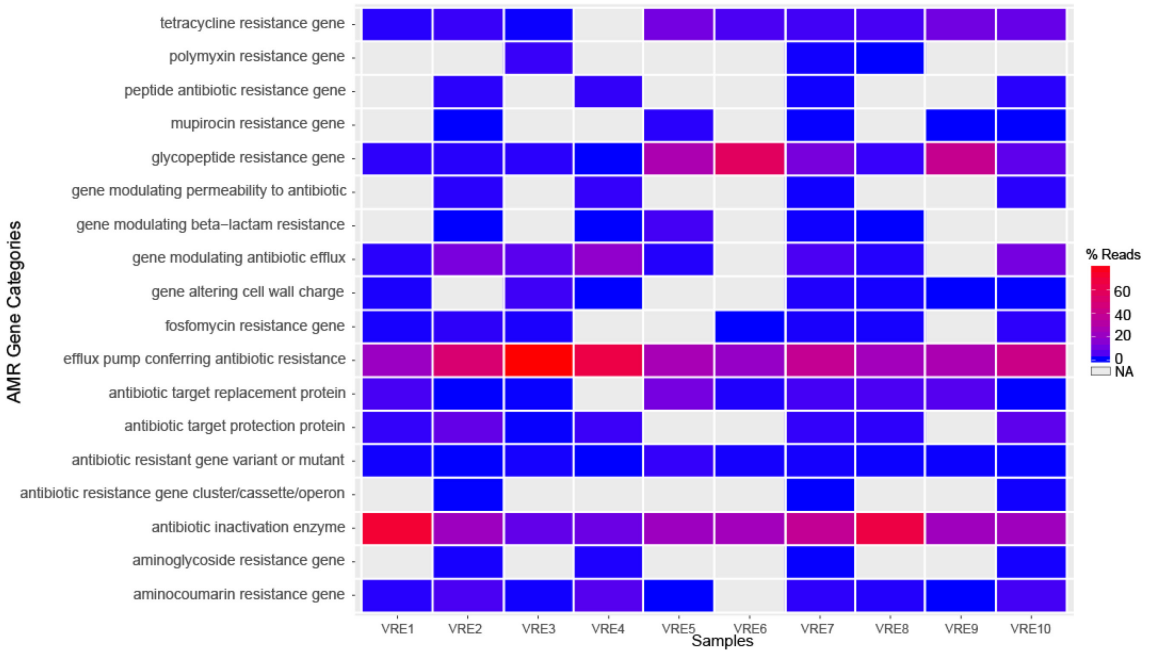

Figure 3: Antibiotic Resistance Detected by CARD. Here, we characterize all of the antimicrobial resistant genes that were detected via sequencing for each sample into the major categories as defined by CARD.

Control Data

We also examined 2 sets of positive controls - one set that had the pure individual organisms in sterile saline (*K. pneumoniae (*KPN*), E. faecalis (*ENFS*),* and *E. faecium (*ENFM*)*), and one set that was comprised of non-VRE/CRO rectal swabs that were spiked in with varying amounts of KPN, ENFS, and ENFM. For our pure individual organisms, when using Kraken, the pure KPN samples corresponded to 86% match to *K. pneumoniae*, with 4% unidentified. ENFM corresponded to an 82% match to *E. faecium*, with 13% unidentified, and ENFS had 96% of its reads correspond to *E. faecalis,* with 1% unidentified. The majority of the other reads were unable to be called down to a species level. Blast results against CARD showed which resistance genes were naturally found in each of the organisms, and aligning back to the reference, we were able to determine where these genes are typically found in these strains.

Our analysis of the spike-in controls is illustrated by the line graph in Figure 4. For the most part, there was no distinguishable difference in spiking in 10^2 organisms compared to no spike ins at all, and typically only a small difference when we spiked in 10^4 organisms. However, when we analyze the 10^6 spike in, there is a noticeable increase in concentration of that particular organism. The only exception is in the sample for ENFM, where the spike-in for 10^2 was significantly higher than our other samples; we are unsure why this occurred.

We also examined the antibiotic resistance genes that were found in each of the control samples, both for the organisms of interest and for our spiked positive controls, using the customized Kraken CARD database.

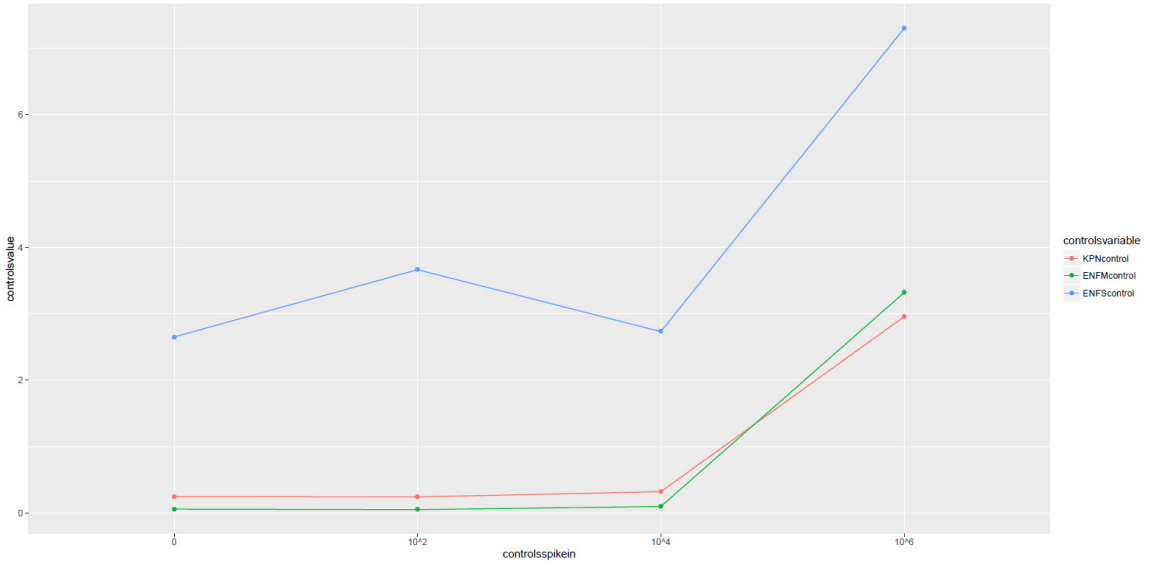

Figure 4: Detection of organism compared to spike in value. For each spike in concentration of each organism, we plotted the percentage of total reads corresponding to the spiked in organism.

Because we uniquely barcoded each of the organisms during sequencing, we can compare the abundance of antibiotic resistance genes present in the bacteria. In *K. pneumoniae*, *KPC* reads only made up about 0.3% of the total number of reads sequenced. Once again, we were unable to classify a specific genotype of *KPC*. *vanA*, which is the vancomycin resistance gene in *E.*

faecium, made up only 0.11% of the total reads sequenced. There were also 5 hits for vanA in *E. faecalis,* but our confidence in these reads is low due to high sequence similarity between *E. faecalis* and *E. faecium*. When looking at *vanB* hits detected in *E. faecium*, we had two hits, supporting this hypothesis. One hit for *vanA* was also seen in *K. pneumoniae*, although that is most likely a false positive due to low count values. *vanB* was the lowest detected antibiotic resistance gene, at only 0.03% of all reads that were sequenced for *E. faecalis*.

For our controls, *KPC* was detected in spiked levels of 10^6 and 10^4, with 89 hits and 10 hits respectively. *vanA* was detected in 10^4 and 10^6 levels, with 3 hits and 421 hits, respectively. The negative control also picked up one hit for *vanA*, which appears to be a false positive. *vanB* was detected in 10^4 and 10^6 spiked levels, with 1 hit and 67 hits respectively.

How well did Nanopore work

Organisms

We wanted to determine effectiveness of the nanopore sequencing to predict taxonomy; with lower capital investment and portability, the MinION has great potential for clinical testing. However, we were concerned that lower yield and accuracy relative to Illumina would prove troublesome for identification. To test this,we sequenced 5 of the VRE-positive rectal swab samples as well as the negative control sample on the Oxford Nanopore MinION. We follow the low-input protocol (see Methods), with each sample sequenced on a R7.3 individual flow cell. We produced an average of 93658 reads per sample (31205 - 187550 range), for an average of 214 Mb sequence (64 - 514Mb range). Our average read length was 3884 with the MinION, as compared to the 2x300 reads generated by our Illumina run. We then applied Kraken to classify the taxonomy of the samples, as previously applied to the Illumina samples.

The nanopore data largely detected the same species as found via the Illumina data, especially for the top three organisms. As with the Illumina data, we plotted the frequency of reads classified as unclassified, viral/fungal, and top bacterial organisms. The negative control had greater than 50% human reads, as seen with the Illumina data, but otherwise mainly comprised of reads aligning to other bacteria. For the 5 VRE positive samples analyzed, human reads took up no more than 20% of all reads, with the majority under 5% human.

Nanopore sequencing provides a critical advantage in rapid discovery and classification of the species in a given sample. As reads are completed by the nanopore sequencer, we can rapidly process them and receive an accurate classification of the samples. We can identify reads from the most prevalent species shortly after sequencing begins, while less frequent species require longer to detect.

Antibiotic Resistance

Nanopore sequencing determined corresponding identified antibiotic resistance genes corresponding to those detected by the Illumina Miseq. We did not detect *vanA*, *vanB*, or *KPC* genes in our negative control. For the VRE positive samples, almost all of the samples were able to detect vanA, the exception being VRE10. However, the number of reads that were detected were extremely low, with none of them above 30 hits, and samples VRE3 and VRE7, for example, only had 2 hits and 1 hit for *vanA,* respectively. The only sample that had *vanB* was sample VRE10, with a total of 4 hits. VRE7 was the only sample that had *KPC* detected, with 227 total hits, with the time of first detection at 2.33 minutes (Figure 6). The number of *vanA* and *vanB* containing reads detected is low, but unsurprising considering these genes are usually found on the chromosome, suggesting low coverage of the full organism itself. *KPC* has a higher

count, but this gene is commonly found on a plasmid, which can have multiple copies depending on the origin of replication for the *K. pneumoniae* plasmid.

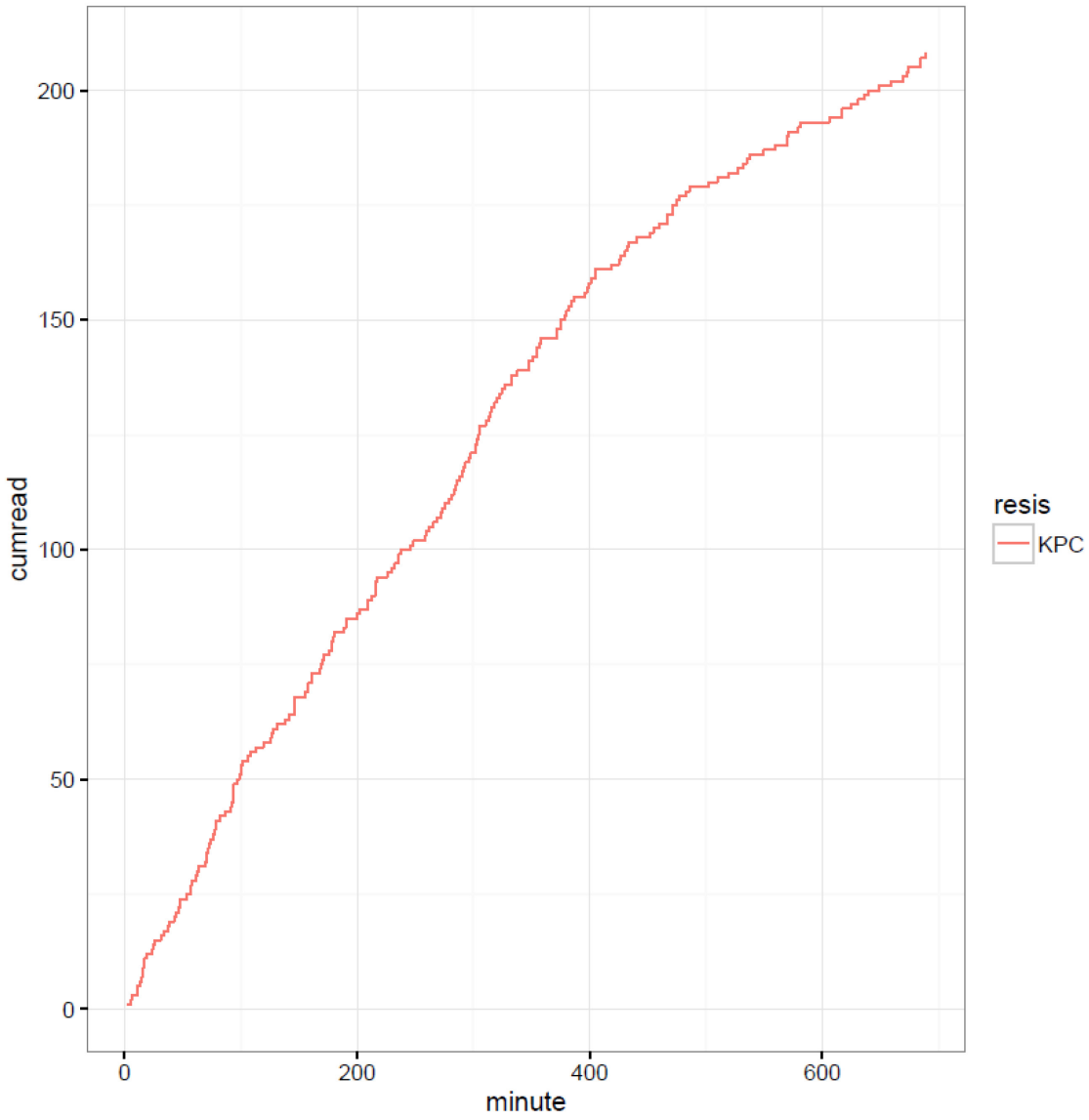

Figure 5: Timeplot of Carbapenem Resistance. Due to the high count of carbapenem-resistant K. pneumonia in VRE7, we examined the time of detection for reads of KPC. The cumulative number of reads of KPC is plotted over time for reads sequenced by the Oxford Nanopore. The first time of detection was 2.33 minutes for KPC

We also used the customized Kraken database for antibiotic resistance developed from CARD on our nanopore sequencing data. Similarly to the Illumina data, we found that the number of reads corresponding to antibiotic resistance were extremely low (less than one percent) in

relation to all of the reads that were sequenced. Many of the reads corresponded to efflux pump genes conferring antibiotic resistance, although there were also high quantities of genes encoding antibiotic inactivation enzymes including glycopeptide resistance genes, and macrolide resistance genes. When we looked at calling specific resistance genes, we found that Nanopore did not detect vancomycin resistance genes in all of our samples. In fact, of the 5 clinical samples that were run on the Nanopore, only two of the samples, 5 and 6, detected the *vanA* gene, at 5 hits and 8 hits respectively. No hits for *vanB* were detected in any of our samples sequenced on the Nanopore. *KPC* was only detected in sample 7, with 202 hits. Like the Illumina data, we were unable to characterize the specific *KPC* genotype. One noteworthy caveat of the Nanopore data is that we had fewer reads overall than from the Illumina data, and due to the comparatively high error rate, it may be difficult to confidently identify variants by nanopore with our current system.

Comparison of NGS to Nanopore

Compare Taxonomy

Using both platforms, we were able to pick up both VRE organisms (*E. faecium, E. faecalis*), and CRO organisms (*K. pneumoniae*) in all of our samples as we saw with culture. When examining the human reads in each sample, we found that the percentage of reads are very similar across both platforms. However, for two samples, 6 and 10, we found a large discrepancy in the percentage of human reads detected (19.5% human on Nanopore vs 0.75% on Illumina for sample 6 and 17.86% on Nanopore vs 0.14% on Illumina for sample 10). ENFS was also similar across both platforms, with the exception being sample 3 (7.69% on Nanopore vs 1.95% on Illumina). ENFM and KPN values were comparable for all samples on both platforms. Other

bacterial organisms present demonstrated variability between the methods as well, which may be partly due to the varying amount of input DNA and bias introduced through PCR steps.

In our hands, Illumina sequencing was superior to nanopore for AMR detection. When utilizing the Kraken CARD database, we were able to detect *vanA* in all of our samples on the Illumina platform, although some of this may be due to sample bleed over and similarities between *vanA* and *vanB*; however, we were unable to detect *vanA* in sample 10 using the Nanopore sequencer. *vanB* was detected on both platforms. When looking at the *KPC* gene, both platforms were able to detect *KPC* in sample 7. While sample 10 had *KPC* detected using the Illumina platform, the number of hits were so low that it may be difficult to determine whether it was a sequencing error or a true hit. Differentiating between different strain types is still difficult on both platforms. With *KPC* genes making up a very small portion of reads, even when looking at the individually sequenced *K. pneumoniae*, it is difficult to determine nucleotide-level resolution of the *KPC* gene are in our sample. Deeper sequencing on our samples is necessary to resolve this issue.

Discussion:

Most studies on the application of NGS as a diagnostic tool from clinical specimens have focused on sterile site specimens such as CSF, tissue and blood. Instead, we wanted to test using a complex specimen source (i.e. rectal swabs) where inhibitors and diverse gastrointestinal microorganisms can complicate sample preparation and data analysis. We have demonstrated the concordance of metagenomic next-generation sequencing with classical culture results, both using Illumina sequencing and nanopore sequencing. These results provide a proof of principle of this application in a clinical setting; not only do we still receive the information from the

original culture based test, but we gain more in depth information on the nature of the organism.

Sequencing provides valuable data on the entire spectrum of the microbiome present in the sample; these insights may suggest specific clinical concerns or treatment courses. Similarly, we gain both the type of antibiotic resistance (innate mutation or acquired gene) and the variety and frequency of resistance present in the sample - tailored therapies (e.g. Augmentin) can then be suggested, avoiding unnecessary escalation of antibiotic use.

Both Illumina and Oxford Nanopore sequencing have the potential to provide speed and automation advantages over more conventional tests. Next-generation sequencing is adaptable into a highly automatable protocol, allowing limited hands-on time and higher lab throughput in principle. The sequencing run itself is often faster than culture tests, with Illumina runs completing in 16 hrs, and nanopore runs providing actionable data within the first few minutes if live streaming data and notification is utilized, which is currently being tested (Cao et al. 2016; Votintseva et al., 2017).

Strain and context information can also provide useful insights in our data. The hyper-accurate (99.9%) Illumina data can identify mutations down to the single nucleotide level, allowing us to identify substrains of the same bacteria or genotypes of an antimicrobial resistance gene. This will allow better tracking of infection sources and resistance gene transfers, working towards prevention of transmission to vulnerable populations. Recognizing strain type is particularly important in the clinical environment for the early identification of microorganism outbreaks and to identify likely sources of transmission in the healthcare environment. While we were unable to specifically characterize genotypes of antimicrobial resistance genes due to low coverage of those genes, long read nanopore sequencing data can provide context for the

location of antimicrobial resistance genes, giving a clear idea of the location of these genes, both the bacteria they are housed in and the physical location in the genome or plasmid carrier. This information can help to develop specific treatments for the bacteria despite their resistance profile.

Cost of sequencing is still a large issue, though one which is being rapidly improved. DNA sequencing costs have been dropping **faster** than Moore's Law for the past 10 years (Wetterstrand 2013); with the advent and addition of nanopore sequencing and the refocus of Illumina's R&D on instruments for clinical applications, it is likely that tests like these will become part of the standard of care in the foreseeable future

Materials and Methods:

Rectal Swabs: We collected 10 remnant rectal Liquid Amies Elution Swabs (Eswab; Copan, Brescia, Italy) previously evaluated for the presence of vancomycin-resistant enterococci (VRE) and carbapenem-resistant Gram-negative organisms (CROs). All 10 swabs were determined to be positive for vancomycin-resistant Enterococci using our standard selective chromogenic culture method (VRESelect™ media, Biorad, Hercules, CA). One of the ten samples (sample #7) was also determined to be positive for a KPC-producing *Klebsiella pneumoniae* using the CDC broth enrichment method and the CheckDirect CPE Screen assay (CheckPoints, Wageningen, The Netherlands) multiplex PCR on the BD MAX instrument (Becton Dickinson, Franklin Lakes, New Jersey).

Controls: 10 rectal Eswab specimens that were found to be negative for both VRE and CROs using standard culture techniques as described above were pooled together, and 500uL of that mixture were included as a negative control. Additionally, three positive controls were included

where aliquots of the pooled negative control was then spiked with varying levels of known ATCC organisms. Three positive controls were spiked with at either 10^2, 10^4 or 10^6 CFU/mL of three ATCC organisms: *K. pneumoniae* ATCC BAA-1705 (bla_{KPC} positive), *Enterococcus faecalis* ATCC 51299 (*vanB* positive) and *Enterococcus faecium* ATCC 700228 (*vanA* positive).

DNA isolation: At the completion of standard of care testing, the swabs were de-identified and were frozen at -70°C until DNA extractions were performed. Total genomic DNA was extracted from 500 µl of the rectal Eswab broth using the Zymo ZR Fungal/Bacterial Miniprep Kit (Zymo, Irvine, CA). Steps were followed as provided with a modification of vortexing by hand at top speed for 5 minutes. Final elution (using provided kit) was 100uL for each sample.

Illumina Sequencing: NexteraXT™ DNA Sample Prep Kit (Illumina, San Diego, CA) was used for DNA fragmentation and library preparation. Briefly, extracted DNA was quantified (ng/µl) on a Qubit instrument to determine total DNA (ng). One nanogram of each sample, diluted down to 0.2ng/uL, was used. This DNA was then tagged and fragmented using the Nextera XT transposome (as provided in the kit). Adapters were then attached to the library and PCR amplified for 12 cycles. Cleanup was performed using 0.6x AMPure. Up to 4 samples at a time were normalized and pooled using bead based normalization as described in the protocol. Each set of pooled samples were then sequenced on a MiSeq v3 2x75 flowcell for up to 16 hours.

Oxford Nanopore Sequencing: The low-input genomic DNA sequencing kit protocol for SQK-MAP006 was used for library prep. 50ng of each sample was first sheared using the Covaris G-tube to 10kb @ 5000rpm for 1 min on each side. End-repair and dA-tailing were then performed using the UltraII End-Prep buffer and End-Prep enzyme mix, with incubation times at 20 and 65C for 5 minutes each, and followed with 1X AMPure cleanup. PCR adapters are then ligated to the sample, followed by PCR of the sample for 18 cycles and 0.4x AMPure cleanup. 1ug of the

cleaned up sample is then taken into End-Prep, incubated at 20C for 5 minutes and 65C for 5 minutes, followed by 1X AMPure cleanup. Adapters and tethers specific for Oxford Nanopore sequencing were then ligated onto the sample, and purified using 50uL MyOne-C1 beads and Bead Binding Buffer (provided in the SQK-MAP006 Sequencing Kit). Libraries were eluted in 25uL elution buffer, and 8uL of the pre-sequencing mix were loaded onto a R7.3 flowcell and sequenced using MinKnow. Sequences were then basecalled using Metrichor. Flowcells were reloaded with another 8uL of pre-sequencing mix after 24 hours.

Bioinformatics: For Nanopore data, we first filtered out the reads based on PHRED quality scores, and for the analyses currently listed, proceeded with the 2d high quality reads. For all samples, taxonomy analysis from the metagenomics sequencing data was performed with Kraken (Wood and Salzberg 2014). Kraken results for microbiome composition results were plotted using Krona for interactive taxonomy visualization (Ondov, Bergman, and Phillippy 2011). We generated a custom kraken database using the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al. 2013) to search for antibiotic resistance genes. BLAST was also used as a more conventional check for specific genes in CARD (specifically *vanA*, *vanB*, and *KPC*). All other plots were generated using R, specifically the ggplot package.

References:

Amberpet, Rajesh, Sujatha Sistla, Subhash Chandra Parija, and Molly Mary Thabah. 2016. "Screening for Intestinal Colonization with Vancomycin Resistant Enterococci and Associated Risk Factors among Patients Admitted to an Adult Intensive Care Unit of a Large Teaching Hospital." *Journal of Clinical and Diagnostic Research: JCDR* 10 (9): DC06–09.

Arias, C. A., G. A. Contreras, and B. E. Murray. 2010. "Management of Multidrug-Resistant Enterococcal Infections." *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 16 (6): 555–62.

Ashton, Philip M., Satheesh Nair, Tim Dallman, Salvatore Rubino, Wolfgang Rabsch, Solomon Mwaigwisya, John Wain, and Justin O'Grady. 2015. "MinION Nanopore Sequencing Identifies the Position and Structure of a Bacterial Antibiotic Resistance Island." *Nature Biotechnology* 33 (3): 296–300.

Berg, Michael G., Deanna Lee, Kelly Coller, Matthew Frankel, Andrew Aronsohn, Kevin Cheng, Kenn Forberg, et al. 2015. "Discovery of a Novel Human Pegivirus in Blood Associated with Hepatitis C Virus Co-Infection." *PLoS Pathogens* 11 (12): e1005325.

Budding, Andries E., Matthijs E. Grasman, Anat Eck, Johannes A. Bogaards, Christina M. J. E. Vandenbroucke-Grauls, Adriaan A. van Bodegraven, and Paul H. M. Savelkoul. 2014. "Rectal Swabs for Analysis of the Intestinal Microbiota." *PloS One* 9 (7): e101344.

Cao, Minh Duc, Devika Ganesamoorthy, Alysha G. Elliott, Huihui Zhang, Matthew A. Cooper, and Lachlan J. M. Coin. 2016. "Streaming Algorithms for Identification of Pathogens and Antibiotic Resistance Potential from Real-Time MinION(TM) Sequencing." *GigaScience* 5 (1): 32.

Eckburg, Paul B., Elisabeth M. Bik, Charles N. Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R. Gill, Karen E. Nelson, and David A. Relman. 2005. "Diversity of the Human Intestinal Microbial Flora." *Science* 308 (5728): 1635–38.

Liu, Yi-Yun, Yang Wang, Timothy R. Walsh, Ling-Xian Yi, Rong Zhang, James Spencer, Yohei Doi, et al. 2016. "Emergence of Plasmid-Mediated Colistin Resistance Mechanism MCR-1 in Animals and Human Beings in China: A Microbiological and Molecular Biological Study." *The Lancet Infectious Diseases* 16 (2). Elsevier: 161–68.

Loman, Nicholas J., Raju V. Misra, Timothy J. Dallman, Chrystala Constantinidou, Saheer E. Gharbia, John Wain, and Mark J. Pallen. 2012. "Performance Comparison of Benchtop High-Throughput Sequencing Platforms." *Nature Biotechnology* 30 (5): 434–39.

McArthur, Andrew G., Nicholas Waglechner, Fazmin Nizam, Austin Yan, Marisa A. Azad, Alison J. Baylay, Kirandeep Bhullar, et al. 2013. "The Comprehensive Antibiotic Resistance Database." *Antimicrobial Agents and Chemotherapy* 57 (7): 3348–57.

Naccache, Samia N., Karl S. Peggs, Frank M. Mattes, Rahul Phadke, Jeremy A. Garson, Paul Grant, Erik Samayoa, et al. 2015. "Diagnosis of Neuroinvasive Astrovirus Infection in an Immunocompromised Adult with Encephalitis by Unbiased next-Generation Sequencing." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 60 (6): 919–23.

Nordmann, Patrice, Thierry Naas, and Laurent Poirel. 2011. "Global Spread of Carbapenemase-Producing Enterobacteriaceae." *Emerging Infectious Diseases* 17 (10): 1791–98.

Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. "Interactive Metagenomic Visualization in a Web Browser." *BMC Bioinformatics* 12 (September): 385.

Samra, Zmira, Orit Ofir, Yinon Lishtzinsky, Liora Madar-Shapiro, and Jihad Bishara. 2007. "Outbreak of Carbapenem-Resistant Klebsiella Pneumoniae Producing KPC-3 in a Tertiary Medical Centre in Israel." *International Journal of Antimicrobial Agents* 30 (6): 525–29.

Simner, Patricia J., Isabella Martin, Belita Opene, Pranita D. Tamma, Karen C. Carroll, and Aaron M. Milstone. 2016. "Evaluation of Multiple Methods for Detection of Gastrointestinal Colonization of Carbapenem-Resistant Organisms from Rectal Swabs." *Journal of Clinical Microbiology* 54 (6): 1664–67.

Snitkin, Evan S., Adrian M. Zelazny, Pamela J. Thomas, Frida Stock, NISC Comparative Sequencing Program Group, David K. Henderson, Tara N. Palmore, and Julia A. Segre. 2012. "Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella Pneumoniae with Whole-Genome Sequencing." *Science Translational Medicine* 4 (148): 148ra116.

Suwantarat, Nuntra, Ava Roberts, Jamie Prestridge, Renee Seeley, Sharon Speser, Christopher Harmon, Chi Zhang, et al. 2014. "Comparison of Five Chromogenic Media for Recovery of Vancomycin-Resistant Enterococci from Fecal Samples." *Journal of Clinical Microbiology* 52 (11): 4039–42.

Timp, W., A. M. Nice, E. M. Nelson, V. Kurz, K. McKelvey, and G. Timp. 2014. "Think Small: Nanopores for Sensing and Synthesis." *IEEE Access* 2: 1396–1408.

Vincent, Caroline, Mark A. Miller, Thaddeus J. Edens, Sudeep Mehrotra, Ken Dewar, and Amee R. Manges. 2016. "Bloom and Bust: Intestinal Microbiota Dynamics in Response to Hospital Exposures and Clostridium Difficile Colonization or Infection." *Microbiome* 4 (March): 12.

Votintseva, A. A., Bradley, P., Pankhurst, L., Del Ojo Elias, C., Loose, M., Nilgiriwala, K., … Iqbal, Z. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole genome sequencing of direct respiratory samples. Journal of Clinical Microbiology. https://doi.org/10.1128/JCM.02483-16

Wetterstrand, K. A. 2013. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." *Human Genome News / National Center for Human Genome Research, National Institutes of Health*.

Wilson, Michael R., Samia N. Naccache, Erik Samayoa, Mark Biagtan, Hiba Bashir, Guixia Yu, Shahriar M. Salamat, et al. 2014. "Actionable Diagnosis of Neuroleptospirosis by next-Generation Sequencing." *The New England Journal of Medicine* 370 (25): 2408–17.

Wood, Derrick E., and Steven L. Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biology* 15 (3): R46.

Chapter 3: A Clinical Case Study Involving an Extensively Drug Resistant Strain of *Klebsiella pneumoniae*

Authors Note:

Part of this work below is adapted from a clinical case study that is set to be submitted for publication in 2017. In the following chapter, I describe some of my work involving this case study, including whole genome sequencing on both the Illumina Miseq and the Oxford Nanopore MinION, genome assembly, and analysis of the genome itself. All of the clinical tests were performed by the Clinical Microbiology department.

Reference to the case study:

Simner, P. J., Antar, A., Hao, S., Gurtowski, J., Tamma, P., Rock, C., Opene, B.N., Tekle, T., Carroll, K., Schatz, M., & Timp, W. (2017). *Acquisition and Mobilization of Resistance Elements in a Patient Infected with a Hypermucoviscous and Extensively Drug-Resistant OXA-181-producing Klebsiella pneumoniae.* Manuscript in Preparation, Johns Hopkins University.

Abstract:

Klebsiella pneumoniae is a growing concern around the world, due to its ability to be acquired in a healthcare setting, as well as its rapid development of antimicrobial resistance, especially to carbapenems.

In this clinical case study, we took 12 isolates that were extracted from a patient coinfected with multiple strains of K. pneumoniae, and sequenced them on an Illumina Miseq. 9 of these isolates were also sequenced on the Oxford Nanopore MinION. A de novo assembly of all the isolates were performed using the SPAdes genome assembler for the Illumina data, and canu for the nanopore data. Annotations were performed using a combination of prokka and blast of various databases, including the Comprehensive Antibiotic Resistance Database (CARD).

We found that there were two distinct strains within the patient, one containing a single plasmid and confirmed to be hypermucovirulent by clinical testing, and another that was extensively drug resistant, containing 6 to 7 plasmids, which harbored multiple antibiotic resistance genes. Using nanopore sequencing, we could assemble the reads into long contigs and characterize samples down to single nucleotide resolution. In addition, we could identify antimicrobial resistance genes as quickly as 15 minutes after starting the sequencing for all of the extensively drug resistant isolates.

Introduction

Infections with *K. pneumoniae* is an increasing problem worldwide, due to drug resistance towards a variety of antibiotics, especially to carbapenems. In the United States, carbapenems are typically not first line treatment, but used when other drugs fail in treatment. *K. pneumoniae* are able to resist carbapenems by producing carbapenemases, which can

hydrolyse carbapenems along with a variety of other antibiotic classes (Munoz-Price et al., 2013). These enzymes were first discovered in 1996 in the United States (Munoz-Price et al., 2013), and genes responsible for KPC enzyme production were first characterized in 2000 (Yigit et al., 2001).

However, other drug-resistant strains of *K. pneumoniae* that do not bear the KPC beta-lactamases also have become increasingly common. A group of these isolates were producing OXA-48 beta-lactamases, which were first discovered from isolates in Turkey (Poirel, Héritier, Tolün, & Nordmann, 2004). These OXA producing K. pneumoniae strains have become increasingly common across the globe, especially in the Middle East and Asia, but have also spread to Europe and North America, including the United States. The OXA producing isolates are especially worrisome as they are found in multi-drug resistant strains, and are resistant to all beta-lactams, which includes carbapenems (Poirel et al., 2004; Poirel, Potron, & Nordmann, 2012). Multiple variants of this gene have also become common, typically differing by a couple of nucleotides, which may impact how resistant the isolate is to various antibiotics (Poirel et al., 2012).

Currently, the gold standard for detecting these particular genes is nucleic acid testing, such as PCR (Kaase, Szabados, Wassill, & Gatermann, 2012). However, differentiating the different gene mutations is not possible without DNA sequencing. With current methods of detection using PCR or hybridization assays, it is difficult to determine what context these genes are in terms of location on the genome or its relation to other genes. In the following case study, we are using whole genome sequencing to examine the genes present with deeper information on sequence, copy number and location to determine what may be the genetic mechanisms behind an extensively drug resistant strain of *K. pneumoniae* that was shown to harbor an oxa-48 like gene.

Case Study Background

A patient was hospitalized at Johns Hopkins Hospital with a *K. pneumoniae* infection. While a variety of antibiotics was administered, none were effective towards treating this infection. Prior to hospitalization at Hopkins, the patient had been visiting India, where the patient had spent a week in the hospital due to a fall. We believe that at some point during this hospitalization abroad, he acquired the *K. pneumoniae* bacteria. Upon return to the United States, the patient was hospitalized, treated for bacteremia and pneumonia, and released. About a month later, the patient was admitted to the Johns Hopkins Hospital. Further testing by both Johns Hopkins Hospital and the Center for Disease Control confirmed that the patient had been infected with an extensively drug resistant strain of *K. pneumoniae* that possessed an OXA-48 like gene.

Materials and Methods

Isolates: Ten isolates were taken from a patient that was being treated for a *K. pneumoniae* infection at Johns Hopkins Hospital over the course of a couple of months. These were taken from numerous locations from the patient's body, including blood, stool, and kidney abscesses, and at various time points throughout his hospital stay. An additional 2 isolates were taken from the patient's hospital room towards the end of his hospital stay (Table 1).

Table 3: Summary of Isolates. This table shows a summary of the different isolates that were collected for this case study, including when it was taken, where, whether or not the isolate was extensively drug resistant, and what sequencing platform each isolate was sequenced on.

Isolate	Hospital Day	Source	Resistance	Sequencer
1	1	Blood	No	Illumina and Nanopore
2	3	Endo/Nasal	No	Illumina and Nanopore
3	8	Sputum	Yes	Illumina Only
4	24	Endo/Nasal	Yes	Illumina and Nanopore
5	32	Kidney Abcess	No	Illumina Only
6	32	Kidney Abcess	No	Illumina and Nanopore
7	39	Kidney Abcess	No	Illumina and Nanopore
8	45	Stool	Yes	Illumina and Nanopore
9	45	Stool	Yes	Illumina and Nanopore
10	56	Blood	Yes	Illumina and Nanopore
11	50	Room	Yes	Illumina Only
12	50	Room	Yes	Illumina and Nanopore

Nucleic Acid Extraction: Genomic DNA was extracted from pure cultures using the MoBio PowerBiofilm DNA Isolation Kit; due to the highly mucoid nature of the isolates, this kit yielded the highest concentration of DNA. One mL of cultured isolate was spun down at 13000g for 1 minute. The supernatant was discarded and the pellet was resuspended in 350uL of provided BF1. This solution was then transferred to a PowerBiofilm Bead Tube. 100uL of BF2 solution was added to the mix, then heated for 5 minutes at 65 degrees C. Tubes were then vortexed horizontally using a MOBIO Vortex Adapter at maximum speed for 10 minutes, centrifuged at 13000g for 1 minute, and supernatant transferred to a new tube. We then added 100uL of BF3 to the supernatant, incubated for 5 minutes at 4 degrees C, and then spun down at 13000g for 1 minute. The resulting supernatant was mixed with 900uL of solution BF4, and then spun down through a spin filter at 13000g for 1 minute. Each spin filter was then washed with 650uL of BF5

and BF6 and spun down at 13000g for 1 minute. Final elution was performed with 100uL of BF7, and quantification was performed using the qubit High Sensitivity kit.

Whole Genome Sequencing: For Illumina sequencing library preparation, we used 1 ng of gDNA with the Nextera XT kit, which applies a tagmentation approach to library preparation. The Illumina libraries were sequenced on the MiSeq using a 2x300 kit, multiplexing up to 4 samples per run, generating 3.6 - 10.7 Million reads per sample (1.99-5.08 Gb yield).

For Nanopore sequencing library preparation, we followed the low input protocol from Oxford Nanopore. Specifically, we sheared 50ng of gDNA in 45uL to 10kbp using Covaris G-tubes at 6000 rpm twice at 1 minute each. We then end-polished and ligated adaptors using the NEB Next Ultra II End-Prep kit, and incubated the samples at 20 degrees C for 5 minutes and 65 C for 5 minutes. Samples were then purified using 1X AMPure XP. Then, we ligated on amplification adaptors from the ONT PCR low-input library preparation kit for 10 minutes at room temperature (SQK-NSK006, R7.3; SQK-LSK208, R9.4). We PCR amplified with LongAmp taq for 18 cycles, then purified with 0.4x AMPure XP. Libraries were checked on the Bioanalyzer for quality control. Around 1.2ug of the cleaned PCR product in 45uL of water was then end prepped using the Ultra II end-prep kit, at 20 degrees C for 5 minutes and 65 degrees C for 5 minutes, then cleaned up using 1X AMPureXP. The sequencing adaptors from ONT containing motor protein were ligated onto the library with NEB Blunt/TA Ligase Master Mix, and incubated for 10 minutes at room temperature. 1uL of tether was added to the mixture, and again incubated for 10 minutes at room temperature. A buffer exchange was performed using MyOne C1 Streptaviadin beads with the ONT provided Bead Bunding Buffer, before final elution in 25uL Elution Buffer for 10 minutes at 37 degrees C. Finally, each isolate was sequenced on an individual nanopore flowcell. For R7.3 flowcells, 150uL of the library loading mix was loaded to the flowcell, consisting of 75uL of 2x Running Buffer, 63uL of nuclease free water, 4uL of fuel

mix, and 8 uL of library. A reload of the flowcell was performed after 24 hours. For R9.4 libraries, 75uL of library loading mix was added to the flowcell, consisting of 37.5uL Running Buffer, 25.5 uL of Library Loading Beads, and 12uL of library. Sequencing on both platforms was performed using MinKNOW, and basecalled using Metrichor 2D basecalling. Runs generated around 41698 - 103587 reads for R7.3 flowcells (0.19-0.5Gb yield), and 189-446 k reads for R9.4 flowcells (0.9-2.5 Gb yield) with an average read length of 4703 kb for R7 runs, and 5418 kb for R9 runs.

FASTQ/Metadata extraction: To extract the sequence and quality data for each read from the fast5 file, we used the poretools fastq /fast5 command from the poretools package (Loman & Quinlan, 2014). Other metadata, such as time, was extracted from the fast5 files using customized python and R scripts.

Genome Assembly: Samples sequenced on the Illumina Miseq were assembled using SPAdes version 3.9.0 using the standard settings (Bankevich et al., 2012). SPAdes works by first performing error correction using Bayes Hammer, which looks at the k-mers present in the reads and corrects first by examining the Hamming distance between the k-mers, followed by Bayesian subclustering (Nikolenko, Korobeynikov, & Alekseyev, 2013). SPAdes then takes these reads to form de Bruijn graphs at various k-mer lengths, and then combines all of these results together to form scaffolds and contigs (Bankevich et al., 2012).

Samples sequenced on the Nanopore were assembled using canu version 1.4. Canu takes the raw reads, counts the k-mers, and then creates overlaps based off of the reads. There are three steps that are performed: 1) read correction, which creates consensus sequences 2) read trimming, to remove sequences that are not high quality, and 3) assembly of the reads into contigs (Koren et al., 2017). To ensure that we capture the shorter, lower coverage plasmids in addition to the full genome, we set options on canu to turn off filtering (using the flag -

contigFilter="2 1000 1.0 1.0 2") and changed the maximum coverage settings to a higher level (-corOutCoverage=1000).

After assembly, contigs were error corrected using Pilon (Walker et al., 2014) of Illumina MiSeq reads aligned to the assembly with bowtie2 (Langmead & Salzberg, 2012). This allows for identification and correction of both single base pair variants, as well as larger structural variants that may have occurred during assembly. To check if contigs needed to be merged together, we used a customized version of Minimus2 (Sommer, Delcher, Salzberg, & Pop, 2007). Briefly, contigs were first aligned to each other with nucmer. If there was sufficient overlap, the proper sequences were cut out using samtools faidx. The relevant sequences were then pieced together using standard Unix tools such as cat and grep.

Genome Annotation: Genomes were annotated using prokka (Seemann, 2014). A custom Klebsiella database was created from the 9 *K. pneumoniae* annotated isolates that were found in GenBank. These strains are 1084, 342, CG43, HS11286, JM45, KCTC_2242, Kp13, MGH_78578, and NTUH_K2044. We annotated for antibiotic resistance genes by both BLASTing using the CARD nucleotide database, and applying RGI_CARD (Jia et al., 2017) to blast versus the protein sequences and ORFs from our assembly.

Phylogeny Analysis: We performed comparisons of strain similarity using the Harvest Tools Suite version 1.1.2. For all of the isolates sequenced on a particular platform, we used parsnp to compare all of the assembled isolates against each other and known reference strains. Parsnp works by first taking the reference strain, and then indexing the sequence using a compressed suffix graph. All of the sequences are then compared to the indexed sequence by aligning to the reference sequence and to each other using MUSCLE. Alignments are then filtered by quality, repetitiveness, and recombination, in order to determine the number of variants, including SNPs, that are in the genome. This SNP data is then used to form a tree and calculate

relatedness between strains. Results were visualized using gingr, which allows for interactive visualization down to the SNP level. (Treangen, Ondov, Koren, & Phillippy, 2014)

MLST/Plasmid Typing: Multilocus sequence typing involves examining a specific set of six or seven housekeeping genes in order to characterize a particular strain of bacteria. We can perform strain identification based on the different allele combinations present within this set of genes. Sequence typing was performed using tools from the Center for Genomic Epidemiology (Lyngby, Denmark). Multi-Locus Sequence Typing (MLST) was performed on each of the assembled isolates using the MLST typing tool (Larsen et al., 2012). Plasmids were identified and typed using PlasmidFinder (Carattoli et al., 2014).

Genome Alignment: Alignment of raw nanopore reads to our assemblies were performed using bwa mem with the --ont2d flag. Resulting bam files were then visualized using IGV viewer. Alignment of assemblies to a known reference strain, HS11286, was performed using the Large Scale Genome Alignment Tool (LASTZ). This is a optimized pairwise aligner based off of BLASTZ, allowing for more user customization, less memory requirements, and increased capabilities to analyze larger sequences (Harris 2007).

Multiple Sequence Alignment (MSA): Multiple sequence alignment allows for comparison of homologous sequences. MSA algorithms typically involve 3 steps: 1) A pairwise aligner that aligns every sequence to each other 2) Creation of a guide tree to weigh the similarities of each of the sequences 3) Progressive alignment based off of the tree that was produced (Thompson, Higgins, & Gibson, 1994). For this case study, we used clustalw to perform our MSA.

Geneious: Some analysis and visualization was performed using Geneious v9.1.0 (http://www.geneious.com, Kearse et al., 2012).

Results

Assembly: Using SPAdes, we assembled each of our genomes with somewhere between 196-719 contigs. Using Nanopore, we assembled each isolate anywhere between 4-29 contigs. After performing LASTZ analysis, we saw full coverage of the K. pneumoniae genome for all of our isolates sequenced on the nanopore, which typically spanned around 5.3-5.4 million base pairs long. For one of our isolates, isolate 1, we got nearly full coverage of our genome in one nice, long contig, while many of the other isolates were broken down into large contigs that spanned several million base pairs.

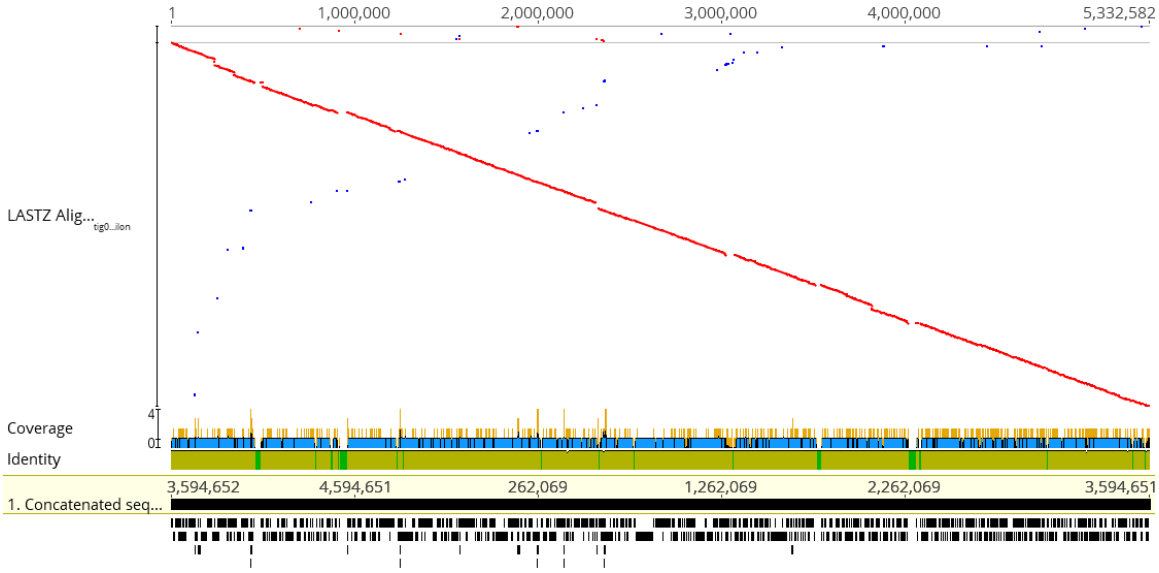

Figure 6: LASTZ Alignment of Contigs Against a Reference K. pneumoniae Genome. Here, we show isolate 1, which nearly had the full genome in one contig, plotted against a reference strain of Klebsiella pneumoniae, HS11286.

Plasmids: For our HMV strains, we found one plasmid for each of the isolates, an IncHI1B plasmid. For our other strains, we found 7 plasmids, which included an IncFIB plasmid, a colKPC plasmid, an IncN2 plasmid, an IncR plasmid, a colpVC plasmid, an IncFII plasmid, and a col(MG828) plasmid. The one exception was isolate 9, where we found six plasmids; there was a fused IncFII/IncR plasmid instead of two distinct plasmids.

Phylogeny comparison: Using both Illumina and Nanopore Sequencing, we were able to compare our assemblies to each of our isolates as well as to known reference strains. We were able to identify two clearly defined groups, the first comprised of isolates 1, 2, 5, 6 and 7, and the second comprised of isolates 3, 4, 8, 9, 10, 11, and 12. This was also shown in the Illumina data as well, where isolates 1, 2, 6 and 7 were found to have very similar phylogenies, and demarcated from isolates 4, 8, 9, 10, and 12. This is similar to clinical results, where the first group of isolates correlated with the isolates determined to be hypermucovirulent (characterized by being extremely sticky), and the second group were determined to be extensively drug resistant (Figure 7). These were also correlated by the differences in SNP's between the strains. When examining the number of SNP's between the isolates of the same strains, there were only few hundred SNP's that differed between them. However, when comparing the two groups, we were able to see a much larger change in differences between the two strains, on the order of three fold greater in magnitude.

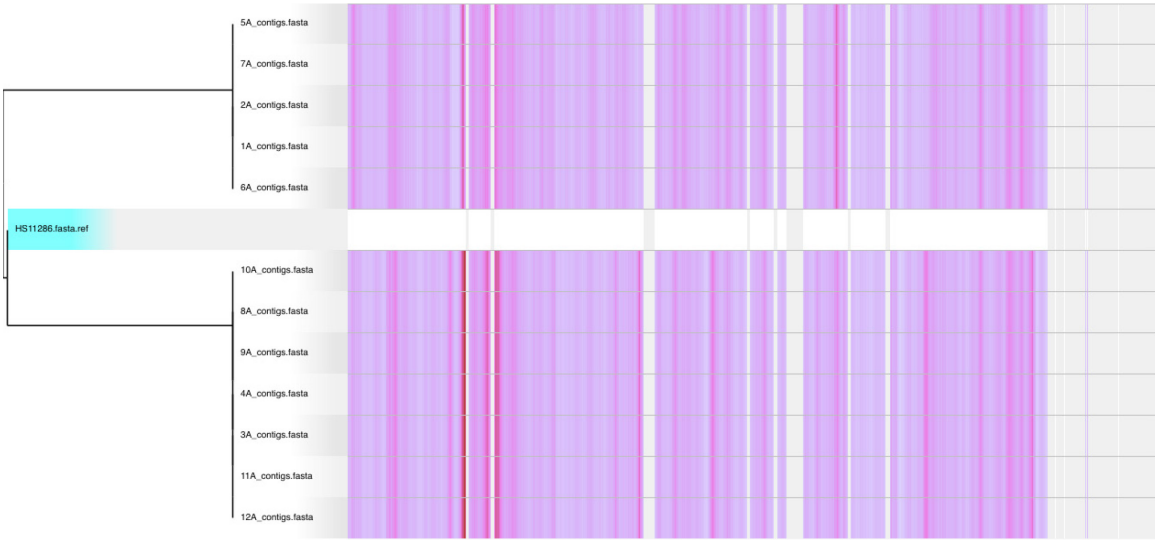

Figure 7: Phylogenic Comparisons of Isolates. Here, we used parsnp to compare the different isolates sequenced on the Illumina Miseq to each other and to a known K. pneumoniae reference genome. You can clearly see that the isolates fall into two separate groups. Results are shown using gingr.

Time to Resistance Detection: Using the Oxford Nanopore MinION, we can examine time of detection of specific genes as results are being sequenced. We were particularly interested in two beta-lactam genes that confer antibiotic resistance, CTX-M-15 and OXA-181. For the five isolates that had these resistance genes, we detected both genes within 15 minutes for all isolates. One of the isolates picked up resistance genes within a minute. The ability to perform this analysis in real time offers a great advantage over other sequencing methods as the information learned can be more quickly used in the management of patient care.

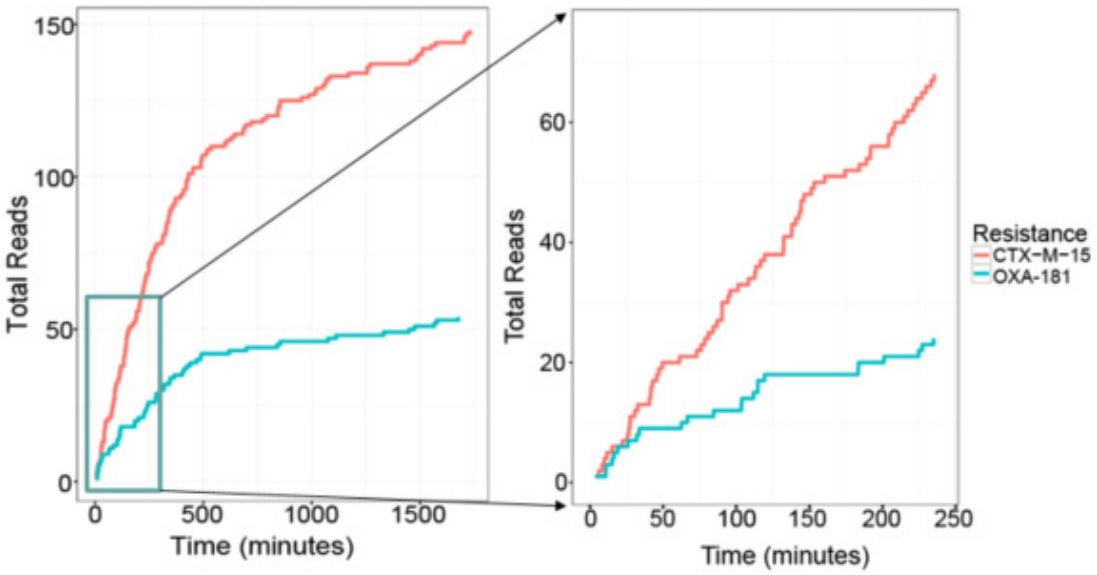

Figure 8: Time of detection of OXA-181 and CTX-M-15. Here, we demonstrate the time of detection for CTX-M-15 and OXA-181 for one isolate using the Oxford Nanopore MinION. The graph on the left shows the span over the first day of sequencing, and the plot on the right shows a zoom in of the first four hours of sequencing (adapted from Simner et al, 2017).

Antibiotic Resistance Analysis: After annotating and blasting, we examined the resistance genes that were found within the isolates. While clinical culture had determined isolates to bear OXA-48 like resistance genes, sequencing results showed that specifically, the isolates bore OXA-181,

a gene similar to OXA-48 with 4 nucleotide substitutions (Poirel et al., 2012). In almost all isolates that harbored OXA-181, we found two copies on a ~13000 base pair colKPC plasmid. In addition, looking at CTX-M-15, we found two copies of the gene, one on a ~180000bp plasmid, and one that has integrated itself into the main chromosome. Other genes of interest that were discovered can be found in the full case report.

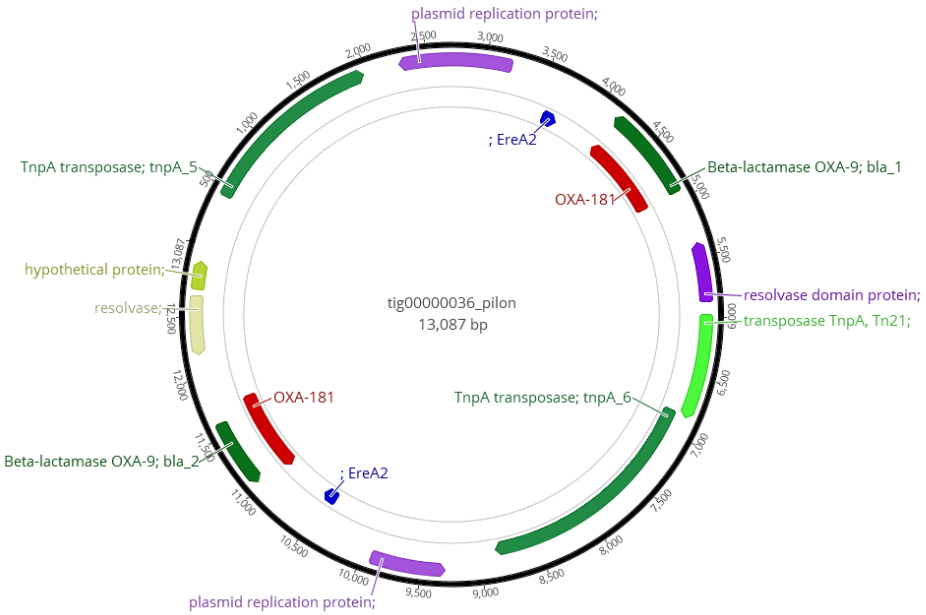

Figure 9: An Annotated OXA plasmid. Two copies of OXA-181 were found on an approximately 13000 basepair long colKPC plasmid. Shown here is the annotated plasmid; the outer circle shows the prokka annotations, including transposable elements and proteins. The inner annotations are the CARD results.

Discussion:

Using the MinION, we could successfully sequence the entire genome of multiple K. pneumoniae isolates using one flow cell per isolate. In addition, we successfully could sequence and identify plasmids present in each of the isolates. We characterized which antibiotic resistance genes were in each isolate, as well as map their proximity to other important features such as mobile elements and replication sites.

By performing whole genome sequencing of the isolate and de novo assembly, we can get a clearer idea as to which genes are present in an isolate. More examination will need to be done to determine how these genes interact with each other and how they may impact antibiotic resistance. For example, in a recent study of K. pneumoniae isolates from the UAE, they found that the OXA-181 gene present in their isolates had actually inserted into the mgrB gene, leading to colistin resistance in those isolates (Sonnevend et al., 2017). Other studies have shown that certain resistance genes are correlated with each other, such as the presence of OXA-like genes with NDM genes, another antibiotic resistance gene that codes for beta-lactam resistance (Ben Nasr et al., 2013). Associated transposable elements are also of great importance, as many of these genes have demonstrated genetic mobility, "jumping" from species to species (van Hoek et al., 2011). Understanding these genetic mechanisms will be of great importance in determining what may lead to the spread of these antibiotic resistance genes.

Ideally, we will one day reach the point where we can perform accurate nanopore-only assemblies. Currently, nanopore-only assembly is achievable through programs such as canu. Error correction, however, is still a major challenge using the nanopore. For this project, initial error correction was performed using Nanopolish, which takes the electrical signal data from the raw reads to correct the consensus assemblies (Loman, Quick, & Simpson, 2015). This can be advantageous, as we can truly get assemblies from start to finish using nanopore-only data. While performing this task, however, we came across two problems: 1) a large amount of time and computing power was necessary to clean up each assembly and 2) at the time of initial analysis, nanopolish was not adapted to analyze R9.4 data. Unfortunately, other viable

nanopore-only correctors do not currently exist. Therefore, we switched over to using Pilon, which utilizes Illumina reads to perform error correction. One of the disadvantages of using Pilon, however, is that multiple rounds of error correction may be necessary in order to correct assemblies. Because each isolate requires a different number of polishes, this process will need to be continually optimized until there are no more changes to be made. It is our hope that return to this method for error correction as future releases of nanopolish become available, and to compare results with those acquired using Pilon.

We can leverage the ability to know the characteristics of the genome with the ability to sequence in real time. For example, once a patient is admitted to the hospital, we could sequence isolates to determine whether or not they are infected with an organism containing a specific antibiotic resistance gene, or look for unique genetic quirks such as the colistin example shown above. This can be used to help personalize the antibiotics that could be administered for more useful treatment, reducing both patient mortality and length of hospital stays. In addition, as sequencing output increases and bioinformatics analysis becomes faster, genome assemblies can also be performed in real time. One such group has already shown that this was possible as a proof of concept by sequencing and assembling a well-established K. pneumoniae strain using the MinION in real time (Cao et al., 2017). This allows us to see the context of the genes as the data is being generated, providing a full picture of the genome structure. We can then also compare this new genome to other known strains to examine the changes from isolates originating around the area and around the world.

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477.

Ben Nasr, A., Decré, D., Compain, F., Genel, N., Barguellil, F., & Arlet, G. (2013). Emergence of NDM-1 in association with OXA-48 in Klebsiella pneumoniae from Tunisia. *Antimicrobial Agents and Chemotherapy*, *57*(8), 4089–4090.

Cao, M. D., Nguyen, S. H., Ganesamoorthy, D., Elliott, A. G., Cooper, M. A., & Coin, L. J. M. (2017). Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nature Communications*, *8*, 14515.

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., … Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, *58*(7), 3895–3903.

Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., … McArthur, A. G. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *45*(D1), D566–D573.

Kaase, M., Szabados, F., Wassill, L., & Gatermann, S. G. (2012). Detection of carbapenemases in Enterobacteriaceae by a commercial multiplex PCR. *Journal of Clinical Microbiology*, *50*(9), 3115–3118.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.*Bioinformatics*, *28*(12), 1647-1649.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. https://doi.org/10.1101/gr.215087.116

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., … Lund, O. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology*, *50*(4), 1355–1361.

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, *12*(8), 733–735.

Loman, N. J., & Quinlan, A. R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* , *30*(23), 3399–3401.

Munoz-Price, L. S., Poirel, L., Bonomo, R. A., Schwaber, M. J., Daikos, G. L., Cormican, M., … Quinn, J. P. (2013). Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. *The Lancet Infectious Diseases*, *13*(9), 785–796.

Nikolenko, S. I., Korobeynikov, A. I., & Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, *14 Suppl 1*, S7.

Poirel, L., Héritier, C., Tolün, V., & Nordmann, P. (2004). Emergence of oxacillinase-mediated resistance to imipenem in Klebsiella pneumoniae. *Antimicrobial Agents and Chemotherapy*, *48*(1), 15–22.

Poirel, L., Potron, A., & Nordmann, P. (2012). OXA-48-like carbapenemases: the phantom menace. *The Journal of Antimicrobial Chemotherapy*, *67*(7), 1597–1606.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , *30*(14), 2068–2069.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., & Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, *8*, 64.

Sonnevend, Á., Ghazawi, A., Hashmey, R., Haidermota, A., Girgis, S., Alfaresi, M., … Pál, T. (2017). Multi-hospital occurrence of pan-resistant Klebsiella pneumoniae ST147 with an ISEcp1-directed blaOXA-181 insertion into the mgrB gene in the United Arab Emirates. *Antimicrobial Agents and Chemotherapy*. https://doi.org/10.1128/AAC.00418-17

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.

Treangen, T. J., Ondov, B. D., Koren, S., & Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, *15*(11), 524.

van Hoek, A. H. A. M., Mevius, D., Guerra, B., Mullany, P., Roberts, A. P., & Aarts, H. J. M. (2011). Acquired antibiotic resistance genes: an overview. *Frontiers in Microbiology*, *2*, 203.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, *9*(11), e112963.

Yigit, H., Queenan, A. M., Anderson, G. J., Domenech-Sanchez, A., Biddle, J. W., Steward, C. D., … Tenover, F. C. (2001). Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae. *Antimicrobial Agents and Chemotherapy*, *45*(4), 1151–1161.

Chapter 4: Influenza Monitoring and Surveillance Using Nanopore Sequencing

Author's Note:

This work was a collaborative effort between the Johns Hopkins Applied Physics Laboratory (APL), our lab, the Center for Excellence in Influenza Research and Surveillance (CEIRS), and the Johns Hopkins School of Public Health, specifically myself, Thomas Mehoke, Yunfan Fan, Rachael Workman, Lauren Sauer, Richard Rothman, Andy Pekosz, Peter Thielen, and Winston Timp. In the following chapter, I will be describing my work on using Nanopore Sequencing to detect and characterize the influenza genome. I would especially like to thank Peter Thielen, Tom Mehoke, and everyone else at APL who provided our lab with samples to sequence and analyze on the Nanopore, as well as performed the Illumina sequencing and aid in the Bioinformatics Analysis, especially for comparison purposes between the Nanopore Sequencer and the Illumina Miseq.

Introduction

Influenza is a single stranded 13.6kb RNA virus that is composed of 8 different segments, ranging in size from approximately 900 base pairs long to 2400 base pairs long. These segments are: PB2, responsible for transcriptase cap binding; PB1, which codes for transcriptase elongation; PA, encoding for protease activity; HA, encoding for Hemagglutinin; NP, which encodes for nucleoproteins, RNA binding, and transport of the viral RNA; NA, which is the neuraminidase segment and also encodes for release of the virus; M1/M2, which encodes for membrane proteins; and NS1/NS2, which encodes for genes such as RNA transport, translation, etc (Life Technologies, 2014)).

Two particular influenza segments are of great interest to researchers. The first is the HA segment, which encodes hemagglutinin, a surface glycoprotein that aids in the binding and fusion of the influenza virus to the cell membrane (Russell et al., 2008). Although there are at least 18 different antigens that are encoded in influenza A viruses, only eight have been known to infect humans, H1, H2, H3, H5, H6, H7, H9, and H10, with the first three, H1, H2, and H3, able to easily spread between humans (Schrauwen & Fouchier, 2014). The second segment of interest is NA, which codes for the neuraminidase glycoprotein, used for keeping new viral particles from sticking to the surface of the cell wall, as well as for helping move the virus towards the target (Shtyrya, Mochalova, & Bovin, 2009). For influenza A viruses, there are eleven different versions of NA. This is reflected in conventional naming of influenza sequences, where we name by hemagglutinin and neuramindase type, e.g. H1N1.

The four major types of influenza are named A, B, C, and D; strains A and B are the ones most dangerous to human health - strain C does not mutate very rapidly and typically only causes mild symptoms in humans (Matsuzaki et al., 2006), and strain D only has been found to infect swine and cattle (Chiara Chiapponi et al., 2016). While there are multiple viral types

associated with influenza A viruses, there are only two major strains of influenza B viruses, typically descending from either B/Yamagata or B/Victoria. This is in part due to how these viruses evolve. All influenza virus types go through something known as antigenic drift, or small changes in the genome that occur as the virus goes through multiple replication cycles. However, influenza A viruses also go through a process known as antigenic shift - an abrupt, major shift that typically leads to new hemagglutinin and/or neuraminidase protein that has never been seen before in circulation due to viral segment rearrangement ("How the Flu Virus Can Change: 'Drift' and 'Shift'| Seasonal Influenza (Flu) | CDC," n.d.; Treanor, 2004).

Because of these mutations, determining what strains of influenza may be circulating each year and how to treat them may be difficult. Typically, influenza outbreaks peak in the Northern Hemisphere between the winter months of December and February. For the 2015-2016 influenza season, the CDC estimates that around 25 million people were afflicted with influenza, with 310,000 of these needing to be hospitalized, and 12,000 people dying, the majority of these being over 65 years of age. These numbers fluctuate depending on how many people are vaccinated with the flu shot each year (current estimations are at 36% vaccination; but estimates indicate an increase of 5% vaccination in the general population would have prevented half a million influenza cases), and how effective the flu shot is against the circulating strains (Rolfes et al. 2016). In addition to the viruses changing from year to year naturally, the influenza virus, in particular H3N2, does not always grow well in eggs during the vaccine development process, making it difficult to produce the flu shot effective each year (Treanor, 2004).

While generally influenza is relatively mild, there can be deadly outbreaks. The most well-known case of this is the "Spanish flu" outbreak in 1918, where in a span of four months, one third of the global population was infected with H1N1, with up to 100 million deaths. Many

of those who were infected and died were young and healthy, unlike typical flu outbreaks (Taubenberger & Morens, 2006). This may have been due to two reasons: 1) the triggering of a cytokine storm, where too many proinflammatory cytokines were released by the immune system, leading to mortality and 2) secondary infections from bacterial pathogens (Morens & Fauci, 2007). Following pandemics did not quite reach the levels of the 1918 epidemic (an H2N2 outbreak in 1957 and a H3N2 outbreak in 1968 led to around 1 million people dying in each epidemic) (Kilbourne, 2006; "Past Pandemics | Pandemic Influenza (Flu) | CDC," n.d.), but there has always been a need to be vigilant, especially . Most recently, a pandemic strain of H1N1 made another appearance in 2009; while fortunately the number of deaths were quite low (only about 20,000 were confirmed by the World Health Organization), up to 200 million people were infected (Garten et al., 2009).

So what makes these strains so deadly? The first is that influenza is becoming more drug resistant to the antivirals available to treat influenza, such as oseltamivir (Tamiflu). This leads to fewer treatment options, especially as this becomes more common (Hayden & de Jong, 2011). In addition, influenza genomes are constantly combining with influenza strains from other species, including swine, human, and avian. This makes infection even more dangerous, since now infections can jump between species. For example, the H2N2 strain responsible for the 1957 outbreak had 3 genes that were from an avian H2N2 strain. The following H3N2 outbreak had two segments that were from an avian H3N2 strain, as well as the neuraminidase segment that was from the previous H2N2 outbreak. (Taubenberger & Morens, 2010). The 2009 pandemic H1N1 outbreak was known as swine flu because the virus contained segments that were a part of swine influenza from North America, Europe, and Asia, as well as human influenza and avian influenza. This is especially alarming due to the fact that avian flu rarely infects humans but can commonly infect pigs. Unfortunately, while many influenza viruses do

transfer cross species now, we still have relatively limited knowledge on influenza that infects other species, such as swine (Schrauwen & Fouchier, 2014).

Influenza Detection

Traditionally, influenza detection in the clinical setting has been performed by first taking a nasal swab from a patient, and then performing diagnostic tests including reverse-transcription PCR, serological tests such as hemagglutination inhibition (HAI) assays, viral cultures, and even commercially available rapid diagnosis tests (Kim & Poudel, 2013). However, this methodology presents problems. For example, serological tests such as HAI requires multiple samples, lots of data, and are very time sensitive. Viral cultures, while accurate, take nearly two weeks to get results and many viruses are challenging to culture. Rapid tests can obtain results in a very short amount of time, but at the cost of sensitivity, at 50-70% sensitive compared other diagnostic techniques. PCR-based assays, on the other hand, appear to alleviate a lot of these problems, and are commonly used in the clinical setting to detect influenza. Reverse-transcriptase PCR (RT-PCR), for example, is more sensitive and specific than previous methods listed above, can be performed relatively quickly in a single tube assay, and is relatively cheap. Even so, this method is relatively limited, as each primer set can only detect one particular influenza strain, and does not provide specific information such as resistance genes and how similar this particular strain is to other circulating strains of the same type (Kim & Poudel, 2013).

Influenza Sequencing

One method to determine resistance gene and strain-specific information is to perform sequencing. Many laboratories use Sanger sequencing; the Center for Disease Control (CDC), for example, has been using Sanger sequencing since the 1980's. However, this method can be laborious, with Sanger library preparation taking up to 2.5 days (Lee, Tang, Kong, & Koay, 2013) This method is relatively manageable with the influenza genome due to its short length, but there is still a lot of room for error, especially if using methods involving plasmid cloning, and is difficult to perform in a high-throughput manner. A simplified Sanger-based sequencing method has been developed for influenza that can cut down the Sanger sequencing down to 8 hours at a more affordable cost, but this still requires primer optimization and lacks the ability to do high-throughput sequencing (Lee et al., 2013).

Within the last 10 years, Next-Generation Sequencing (NGS), has risen in popularity due to its ability to perform high-throughput sequencing, as well as cheaper costs. One of the earliest sequencing runs of influenza was sequencing positive nasal swabs on the Solexa platform. Using a shotgun sequencing approach of clinical samples, 8.39 - 44.56% of reads were identified as influenza, which is "sufficient for viral subtype identification" (Yongfeng et al., 2011). Since then, library preparations for influenza sequencing have been optimized extensively, such as modifying the RT-PCR to get full length sequences with even coverage of all segments using modified primers. Additionally, sequencing has been shown to be very successful at a high-throughput scale (Lee, Lee, Tang, Loh, & Koay, 2016; McGinnis, Laplante, Shudt, & George, 2016). One of the first large-scale sequencing of influenza was performed in 2005, providing full genomes for 209 H3N2 isolates (Ghedin et al., 2005). Today, there are over 16,000 full influenza genomes available in Genbank ("Influenza Genome Sequencing Project | NIH: National Institute of Allergy and Infectious Diseases," n.d.), and that number is continually growing. However, this method is also not without limitations. Unlike Sanger, which could get

reads up to 700 basepairs long, standard NGS techniques provide much shorter reads, typically up to 300 base pairs long. In addition, the error rates for NGS when sequencing influenza on Illumina were noticed to be 0.12% for mismatches, and up to 0.004% for insertions and deletions (Archer et al., 2012).

More recently, third-generation sequencers have been shown to be effective at viral sequencing and subtyping. The first major study to have sequenced influenza on the MinION was a group from New Zealand in 2015. They took one strain, influenza A/New Zealand/316/2014 (H3N2), and sequenced it using three different methods - Sanger sequencing, Illumina Miseq sequencing with library prep from the Illumina Truseq Nano library kit, and the Oxford Nanopore MinION, using the SQK_MAP004 sequencing kit and a R7.3 flowcell. They were able to find that there was a 99% consensus using the Nanopore compared to both Sanger and Illumina sequencing, and were able to obtain full-length reads for all segments with relatively even coverage, over the course of four hours (Wang, Moore, Deng, Eccles, & Hall, 2015). Since then, other groups, such as the Ghedin lab at New York University, have been able to successfully validate the ability for Nanopore to sequence all segments of the influenza genome with full length reads. Other viruses have been sequenced extensively on the MinION, including Ebola and Zika (Faria et al., 2016; Quick et al., 2016).

For this study, we hope to validate the performance of the MinION for sequencing influenza in order to characterize circulating influenza strains and provide a tool that can aid clinicians in providing more rapid, accurate treatment at influenza outbreak sites.

Materials and Methods

Samples and Nucleic Acid Extraction

Clinical nasal swabs testing positive for influenza from the 2015-2016 winter season from the Centers of Excellence for Influenza Research and Surveillance (CEIRS) network at Johns Hopkins University were obtained. In addition, lab grown strains of H1N1 and H3N2, as well as modified genomes for each influenza type consisting of a recombinant N2 segment that doubled the length of that particular segment, were used as positive controls. RNA extraction from clinical samples and lab-grown influenza strains was performed using the MagMax Viral RNA extraction kit, and eluted in 50uL elution buffer. Samples were amplified using multi-segmented PCR to amplify the influenza genome, which utilizes conserved regions on each influenza genome segment for amplification.

Sequencing (Illumina)

To prepare samples for Illumina sequencing, samples were first converted into cDNA using reverse-transcriptase PCR (RT-PCR). One nanogram of each sample was used to prepare libraries for sequencing using a transposon based library prep kit, Nextera XT. 50-80 samples were multiplexed and sequenced on an Illumina NextSeq using a 2x150bp cartridge for 25-30 hours.

Sequencing (Nanopore)

R7.3

As a proof of concept, we first sequenced each of the lab strains individually on a flowcell, using the low input expansion pack protocol for genomic DNA. Samples were first amplified by PCR using ONT-specific tailed primers. One hundred nanograms of amplified sample was then prepared for adapter ligation using the Ultra II End-Prep reaction mix from

NEB, and incubated for 5 minutes at 20C and 5 min at 65C. End-prepped product was then cleaned up using 1X Ampure. Adapter ligation was performed using Nanopore specific adapters and blunt/TA ligation master mix, and incubated for 10 minutes at room temperature. One microliter of ONT tether was added and incubated for 10 additional minutes at room temperature. Samples were then purified using 1x MyOneC1 Streptaviadin beads that was buffer exchanged with the kit provided Bead Binding Buffer. Elution was performed in 15uL of ONT provided elution buffer at 37C for 10 minutes.

Libraries were prepared for MinION loading by combining 6uL of the prepared sample combined with 75uL of RNB, 65uL of nuclease free water, and 4uL of Fuel Mix. For each sample, one R7.3 flowcell was primed twice with a mixture of Running Buffer and fuel mix. 150uL of the prepared library was loaded onto the flowcell. Samples were then sequenced using MinKnow for 48 hours, and basecalled using Metrichor. The flow cell was reloaded with this same mixture after 24 hours of sequencing.

We performed clinical sample sequencing, using the low-input native barcoding protocol. Six H3N2 clinical samples were first amplified by PCR using ONT-specific tailed primers. End repair was performed as detailed above. A unique barcode was ligated onto 0.2pmol of each purified sample with Blunt/TA Ligase Master Mix, incubated for 10 minutes at room temperature, PCR'd with LongAmp Taq for 12 cycles, then purified with 1X Ampure. Sample was then pooled evenly to 700ng DNA total in 58 uL of water. ONT specific adapters and hairpin ligation, elution and library loading proceeded as detailed above, and basecalling was performed on Metrichor using a 2D barcoded workflow.

R9

To determine effectiveness of the R9 iteration of sequencing chemistry for influenza sequencing, we prepared another 9 clinical samples using a low-input barcoding protocol, which allowed us to use samples as low as 20ng. Samples were first amplified by PCR using ONT-specific tailed primers. 0.5nM of each sample was then used for PCR barcoding and purification. QC on each of the samples was performed through quantification using the Qubit, as well as running a BioAnalyzer High Sensitivity Chip in order to determine that samples were of the correct size and that there was no primer/adapter dimer within the sample. Barcoded samples were then evenly pooled together to form 1.2ug of genomic material. We then followed the Oxford Nanopore SQK-NSK007 sequencing kit protocol; although chemistry of proprietary nanopore reagents had been updated, our library preparation steps were unaltered from those described above. Samples were then sequenced on a R9 flowcell for 48 hours using MINKnow. Basecalling was performed on Metrichor using a 2D barcoded RNN basecaller workflow.

R9.4

As Oxford Nanopore began another series of chemistry updates and flowcell iterations, we once more sought to validate efficacy of influenza multiplexing and sequencing on this platform. Two additional sequencing runs were performed using the SQK-LSK108 library prep kit, which is a ligation based kit for adapter ligation. All samples were first amplified by PCR with ONT specific tailed primers. For the first set of samples, we followed the 1D PCR barcoding genomic DNA protocol. Twelve samples were individually end-prepped using the NEB Ultra II End-Prep kit at 20C for 5 minutes and 65 C for 5 minutes, followed by an 1X Ampure cleanup. A unique barcode was then ligated onto 0.2pmol of each sample using blunt TA/ligase master mix, followed by PCR for 12 cycles using LongAmp 2x Master Mix. Samples were purified with 1X

Ampure and evenly pooled together to 1.2ug. The second run consisted of 53 samples barcoded together using the 1D PCR barcoding (96) amplicons protocol. Samples from the first PCR step were then taken into a barcoding PCR step, and then purified using 1x AMPure and pooled together to 1.3ug. Both sets of samples were then end-prepped under the same conditions as before, and cleaned with 1X Ampure. ONT-specific adapters were ligated onto the samples using Blunt/TA Ligase Master Mix for 10 minutes at room temperature. A second cleanup was performed using 0.4X Ampure for the first set of samples, and 0.7X Ampure for the second set of samples. Samples were eluted in 15uL of ONT elution buffer at room temperature for 10 minutes. 12uL of the eluted library was then mixed with 35 uL running buffer, 2.5uL water, and 25.5uL library loading beads to form the library loading mix.

For each set, one R9.4 Spot On flowcell was primed twice with a mixture of running buffer and water, adding 800uL of the mixture into the priming port, and an additional 200uL after five minutes. The library loading mix was then added in droplet fashion into the sample port. Samples were then sequenced for 48 hours using MinKnow. The first run was basecalled on Metrichor using the 1D basecalling + barcoding (450 Mbs) protocol. The second set was basecalled using a command-line version of Albacore.

Bioinformatics Analysis

For nanopore data, samples were aligned to either the H1N1 Giessen 2009 strain (A/Giessen/6/2009(H1N1)), or to the H3N2 Victoria 2011 strain (A/Victoria/361/2011(H3N2)) found from the Open Flu Database (Liechti et al., 2010) using bwa-mem with a specific flag for nanopore reads (--ont2d). Alignment, consensus sequences, and VCF files (to examine SNPs) were generated using samtools mpileup and bcftools (Li et al., 2009). Count and length plots for each sample were generated using R.

To determine similarity of our clinical strains to known strains, we employed two methods. First, we characterized our samples with a customized Kraken database that has over 2000+ strains of influenza from various locations and timepoints. Second, to compare phylogeny, full consensus sequences for each segment from Nanopore and Illumina were then placed through NextFlu (Neher & Bedford, 2015) in order to determine how similar strains were to each other and to other known influenza viruses.

Results

For our R7 runs, our 2D yields ranged from about 4333 reads up to 64000 reads, or approximately 5-85 Mb of data per flowcell. For our R9 run, we generated 142525 2D reads, or around 166.5Mb of data on a single flowcell. For our first 9.4 run, we produced 496428 reads, or about 437.8Mb of data. When comparing to our reference genome, these samples also had high alignment rate against a reference genome - for our first 9.4 run, all of our samples had anywhere between a 70-99.9% alignment rate, with the exception of one sample, which only had a 0.5% alignment rate due to nonspecific amplification during sample preparation.

With almost all of our samples for all chemistries, we obtained full length reads for each segment of the influenza genome (Figure 10). However, we have found a low percentage in number of full length reads especially in the PB2, PB1/PB1-F2, and PA/PA-X segments (generally less than 15% of all reads were full length compared to >70% reads being full length in HA, NP, NA, M1/M2, and NS1/NS2 segments) for a portion of our samples. This appears to be due to mispriming events within the genetic sequence of that segment, and is highlighted in figure 11 and table 5.

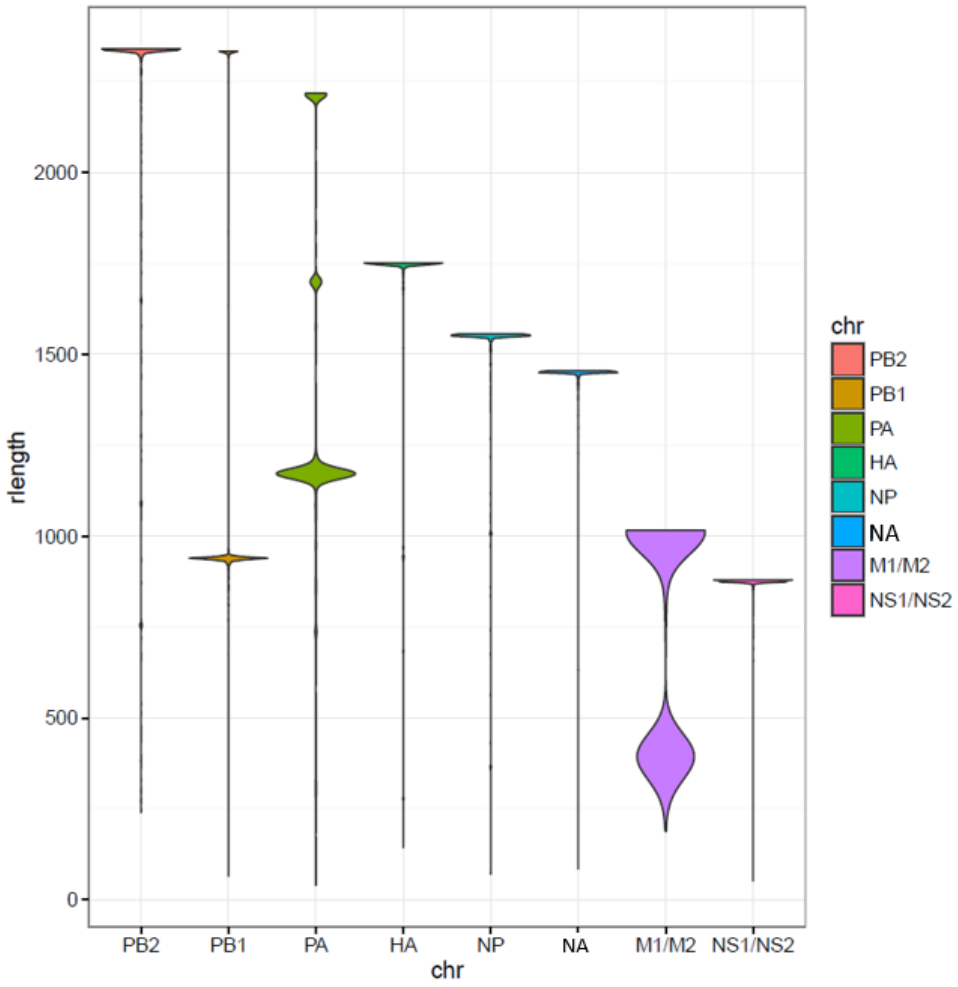

Figure 10: A violin plot showing distribution of read length for each segment. For each segment of the influenza genome, distributions of reads at each length are plotted. The more reads that occurred at that particular length, the wider the plot. For this particular sample, we were able to get full length reads in all of the segments. However, for PB1, PA, and M1/M2, we also saw a lot of shorter fragments, most likely due to the presence of defective interfering particles

Table 4: Number and percent of full length reads for 8 samples: Number of reads that were within one percent of a segment's expected length are given for each sample with percentage of reads that were full length in parenthesis

	BC02	BC03	BC04	BC05	BC06	BC08	BC09	BC11
PB2	4 (2.3)	7 (9.7)	0 (0.0)	43 (75.4)	410 (76.9)	94 (10.1)	48 (28.2)	88 (21.7)
PB1/PB1-F2	11 (2.1)	53 (12.3)	0 (0.0)	38 (26.8)	188 (14.6)	70 (6.1)	30 (9.9)	58 (4.6)
PA/PA-X	67 (8.2)	170 (25.6)	21 (7.1)	197 (67.2)	170 (13.9)	59 (4.6)	56 (8.2)	233 (3.2)
HA	321 (82.7)	317 (89.0)	148 (86.5)	288 (91.7)	1108 (85.4)	545 (47.0)	13 (0.8)	590 (87.7)
NP	141 (68.4)	255 (86.4)	148 (84.6)	552 (85.4)	472 (85.0)	147 (24.1)	68 (26.9)	675 (87.0)
NA	99 (81.8)	376 (85.6)	141 (88.1)	645 (89.0)	1611 (86.4)	475 (68.4)	558 (63.2)	1307 (88.7)
M1/M2	2925 (73.2)	1204 (85.1)	3584 (78.8)	3002 (86.0)	1043 (51.4)	2543 (51.4)	1790 (46.5)	2329 (77.0)
NS1/NS2	2496 (88.7)	1254 (88.1)	2960 (88.9)	2186 (88.7)	3635 (88.0)	0 (0.0)	0 (0.0)	3395 (85.7)

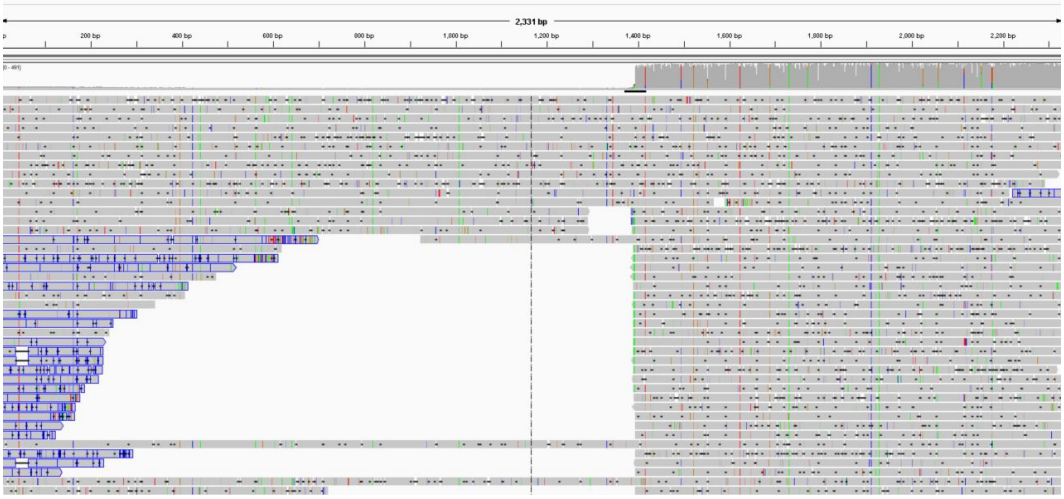

Figure 11: IGV plot showing sequencing coverage for a clinical sample for PB1 against a reference genome. The region in the middle had much lower coverage, probably due to a mispriming event in PB1.

For each of the clinical samples that were run on multiple platforms, we compared the results of multi-segmented PCR and RNA sequencing from the Illumina Miseq to Nanopore sequencing results (on both R7.3 and R9, if applicable) to known influenza strains using NextFlu (Figure 4). We found that all of our sequences were able to be grouped with the correct viral type (for example, with the H3N2 clinical samples that were run on the R9 flowcell, they all were correctly identified as influenza A H3N2 virus). In addition, we obtained correct identifications for all sequencing methods. Occasionally, the R7 influenza sequences were sufficiently different that they were not fully grouped with the other three methods of sequencing; however, they still clustered together with other sequencing methods of the same sample. By using NextFlu, we were also able to see how closely related these strains were related to other Clade 3C influenza strains, and more specifically, see that they were all classified correctly as clade 3C.2a viruses.

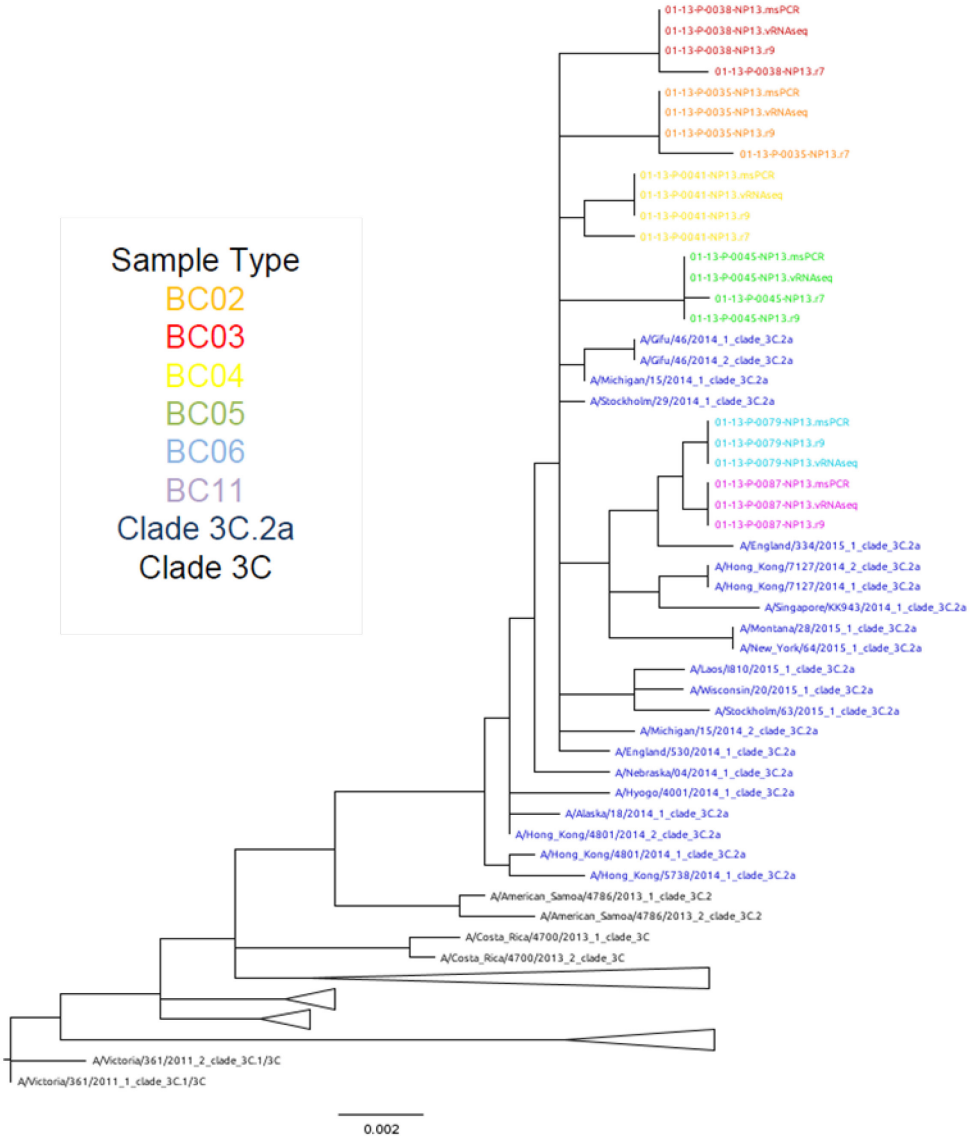

Figure 12: Phylogenic tree comparing our clinical samples (run on Miseq (msPCR and RNAseq) and Nanopore (R7 and R9 versions)) to other strains of influenza. Blue and black samples are strains of clade 3C.2a and 3C, respectively.

We examined the Single Nucleotide Polymorphisms (SNPs) from the Nanopore consensus sequence to the original reference genome. For nanopore, a SNP was called if there was a high percentage of mutation frequency from the raw reads. On the Illumina Miseq platform, a SNP was called if the mutation frequency was at almost 100%. When comparing the mutation frequency on both the Illumina and Nanopore platforms, we found that the Miseq

data gave very clear peaks as to locations where there was a SNP (100% mutation frequency), versus no SNP (no mutation frequency was detected). The nanopore mutation frequency data was expectedly noisier in comparison.In general, mutation frequency tended to range anywhere between 0 and 40% for a single base pair with the R7.3 data. When comparing R7.3 data to R9 data, we found that overall, R9 data, while still noisy, was much improved (Figure 5). In addition, the false positive SNP calling rate dropped when switching from the R7.3 to the R9. Even with these improvements, the Nanopore still is not quite as accurate in calling SNPs compared to the Illumina platform - when comparing the R9 data to the Illumina Miseq, we found that Nanopore tended to call fewer correct SNPs than Illumina did. In addition, Nanopore called up to 5 SNPs per segment that were not identified using the Miseq, leading to a higher false positive rate (Table 2)

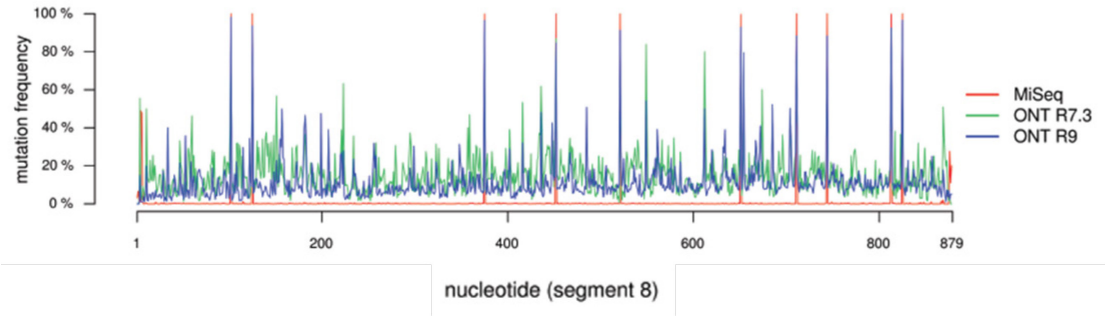

Figure 13: Comparison of SNP calling between the Illumina Miseq and the Oxford Nanopore. Per-base sequencing error rates are higher from the Oxford platform, but penetrant SNPs are distinguishable from sequencing noise for consensus sequence generation.

Table 5: Mutations chart comparing number of SNP's to reference genomes on both Nanopore and Illumina platforms. First value is number of SNP's detected on Nanopore that matched Illumina, second value is number of SNP's detected by Illumina, and value in parenthesis is the number of SNP's called by Nanopore but not Illumina.

	BC02	BC03	BC04	BC05	BC06	BC08	BC09	BC11
PB2	24/30 (5)	24/28 (1)	10/21 (0)	28/28 (1)	28/28 (1)	1/1 (2)	0/0 (0)	28/28 (2)
PB1/PB1-F2	17/18 (3)	17/18 (2)	8/19 (5)	21/21 (3)	23/23 (4)	0/0 (3)	0/0 (1)	21/22 (5)
PA/PA-X	31/32 (4)	27/28 (1)	27/29 (2)	23/24 (1)	25/26 (4)	0/0 (2)	0/0 (1)	25/26 (1)
HA	30/30 (2)	30/30 (1)	29/29 (2)	30/31 (2)	32/32 (0)	4/4 (1)	1/1 (4)	32/32 (1)
NP	15/15 (2)	12/12 (3)	12/13 (4)	14/14 (2)	14/14 (3)	0/0 (1)	0/1 (1)	13/13 (3)
NA	14/14 (1)	15/15 (2)	15/15 (2)	14/14 (1)	17/17 (1)	0/0 (1)	0/0 (2)	16/17 (2)
M1/M2	4/4 (1)	3/3 (0)	3/3 (1)	2/2 (0)	2/2 (1)	0/0 (1)	0/0 (2)	2/2 (1)
NS1/NEP	10/10 (1)	9/9 (2)	8/8 (2)	8/8 (1)	9/9 (3)	2/2 (2)	2/2 (2)	9/9 (2)

Conclusions and Discussion

Using Nanopore sequencing, we were able to characterize clinical influenza samples by examining type and variations within the samples. Full length reads were achievable in the majority of samples tested. We had difficulty obtaining full length reads from PB2, PB1, and PA, which is consistent with other influenza sequencing studies (Saira et al., 2013). These shortened lengths are most likely due to mispriming events. Our sequences aligned well to the known influenza strains; the exception being a sample co-infected with influenza B, which had also amplified with the primers that were used during initial amplification.

We found our sequencing results to be relatively comparable to the Miseq, although Nanopore does not always call as many SNPs as Illumina does, most likely due to a high error rate. This may be improved with adding an error correction step, such as using nanopolish. However, as the MinION technology advances, we saw improvements in results, including yield, number of full length reads, consensus, and accurate SNP calling.

These factors, combined with the MinION's portability and low cost, show promising capabilities for use in monitoring and surveying influenza outbreak sites. Ultimately, we are interested in integrating the nanopore sequencer as a part of a mobile genomics "lab in a suitcase", which can be taken directly to locations of high influenza activity to collect data on site. This will allow us to provide more timely and effective treatment to those who have influenza-like symptoms, as well as examine how circulating strains are being transmitted and evolving over time.

References

Archer, J., Baillie, G., Watson, S. J., Kellam, P., Rambaut, A., & Robertson, D. L. (2012). Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. *BMC Bioinformatics*, *13*, 47.

Chiara Chiapponi, Silvia Faccini, Aurora De Mattia, Laura Baioni, Ilaria Barbieri, Carlo Rosignoli, … Emanuela Foni. (2016). Detection of Influenza D Virus among Swine and Cattle, Italy. *Emerging Infectious Disease Journal*, *22*(2), 352.

Faria, N. R., Sabino, E. C., Nunes, M. R. T., Alcantara, L. C. J., Loman, N. J., & Pybus, O. G. (2016). Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, *8*(1), 97.

Garten, R. J., Davis, C. T., Russell, C. A., Shu, B., Lindstrom, S., Balish, A., … Cox, N. J. (2009). Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science*, *325*(5937), 197–201.

Ghedin, E., Sengamalay, N. A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., … Salzberg, S. L. (2005). Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, *437*(7062), 1162–1166.

Hayden, F. G., & de Jong, M. D. (2011). Emerging influenza antiviral resistance threats. *The Journal of Infectious Diseases*, *203*(1), 6–10.

How the Flu Virus Can Change: "Drift" and "Shift"| Seasonal Influenza (Flu) | CDC. (n.d.). Retrieved April 20, 2017, from https://www.cdc.gov/flu/about/viruses/change.htm

Influenza Genome Sequencing Project | NIH: National Institute of Allergy and Infectious Diseases. (n.d.). Retrieved April 22, 2017, from https://www.niaid.nih.gov/research/genome-sequencing-centers

Kilbourne, E. D. (2006). Influenza pandemics of the 20th century. *Emerging Infectious Diseases*, *12*(1), 9–14.

Kim, D.-K., & Poudel, B. (2013). Tools to detect influenza virus. *Yonsei Medical Journal*, *54*(3), 560–566.

Lee, H. K., Lee, C. K., Tang, J. W.-T., Loh, T. P., & Koay, E. S.-C. (2016). Contamination-controlled high-throughput whole genome sequencing for influenza A viruses using the MiSeq sequencer. *Scientific Reports*, *6*, 33318.

Lee, H. K., Tang, J. W.-T., Kong, D. H.-L., & Koay, E. S.-C. (2013). Simplified large-scale Sanger genome sequencing for influenza A/H3N2 virus. *PloS One*, *8*(5), e64785.

Liechti, R., Gleizes, A., Kuznetsov, D., Bougueleret, L., Le Mercier, P., Bairoch, A., & Xenarios, I. (2010). OpenFluDB, a database for human and animal influenza virus. *Database: The Journal of Biological Databases and Curation*, *2010*, baq004.

Life Technologies. (2014). *INFLUENZA A WHOLE-GENOME SEQUENCING*. Retrieved from https://tools.thermofisher.com/content/sfs/brochures/Influenza_A_Typing_App_Note.pdf

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Matsuzaki, Y., Katsushima, N., Nagai, Y., Shoji, M., Itagaki, T., Sakamoto, M., … Nishimura, H. (2006). Clinical features of influenza C virus infection in children. *The Journal of Infectious Diseases*, *193*(9), 1229–1235.

McGinnis, J., Laplante, J., Shudt, M., & George, K. S. (2016). Next generation sequencing for whole genome analysis and surveillance of influenza A viruses. *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology*, *79*, 44–50.

Morens, D. M., & Fauci, A. S. (2007). The 1918 influenza pandemic: insights for the 21st century. *The Journal of Infectious Diseases*, *195*(7), 1018–1028.

Neher, R. A., & Bedford, T. (2015). nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* , *31*(21), 3546–3548.

Past Pandemics | Pandemic Influenza (Flu) | CDC. (n.d.). Retrieved April 20, 2017, from https://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., … Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature*, *530*(7589), 228–232.

Rolfes MA, Foppa IM, Garg S, Flannery B, Brammer L, Singleton JA, et al. Estimated Influenza Illnesses, Medical Visits, Hospitalizations, and Deaths Averted by Vaccination in the United States. 2016 Dec 9; 2017 Apr 23; https://www.cdc.gov/flu/about/disease/2015-16.htm

Russell, R. J., Kerry, P. S., Stevens, D. J., Steinhauer, D. A., Martin, S. R., Gamblin, S. J., & Skehel, J. J. (2008). Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 17736–17741.

Saira, K., Lin, X., DePasse, J. V., Halpin, R., Twaddle, A., Stockwell, T., … INSIGHT FLU003 Study Group. (2013). Sequence analysis of in vivo defective interfering-like RNA of influenza A H1N1 pandemic virus. *Journal of Virology*, *87*(14), 8064–8074.

Schrauwen, E. J., & Fouchier, R. A. (2014). Host adaptation and transmission of influenza A viruses in mammals. *Emerging Microbes & Infections*, *3*(2), e9.

Shtyrya, Y. A., Mochalova, L. V., & Bovin, N. V. (2009). Influenza virus neuraminidase: structure and function. *Acta Naturae*, *1*(2), 26–32.

Taubenberger, J. K., & Morens, D. M. (2006). 1918 Influenza: the mother of all pandemics. *Emerging Infectious Diseases*, *12*(1), 15–22.

Taubenberger, J. K., & Morens, D. M. (2010). Influenza: the once and future pandemic. *Public Health Reports*, *125 Suppl 3*, 16–26.

Treanor, J. (2004). Influenza Vaccine — Outmaneuvering Antigenic Shift and Drift. *The New England Journal of Medicine*, *350*(3), 218–220.

Wang, J., Moore, N. E., Deng, Y.-M., Eccles, D. A., & Hall, R. J. (2015). MinION nanopore sequencing of an influenza genome. *Frontiers in Microbiology*, *6*, 766.

Yongfeng, H., Fan, Y., Jie, D., Jian, Y., Ting, Z., Lilian, S., & Jin, Q. (2011). Direct pathogen detection from swab samples using a new high-throughput sequencing technology. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *17*(2), 241–244.

Chapter 5: Overall Discussions and Conclusions

Discussion and Conclusions

We successfully detected pathogenic organisms of interest when performing metagenomic shotgun sequencing, both on the Illumina Miseq and Oxford Nanopore platforms. In particular, from the peri-rectal shotgun sequencing experiment, we could easily detect vancomycin resistant Enterococci, such as *E. faecalis* and *E. faecium*, as well as carbapenem resistant organisms, such as *K. pneumoniae*. Specific antimicrobial resistance genes, such as *vanA* and *vanB*, were more difficult to detect, suggesting that both platforms require more depth of coverage per sample to glean insights into specific gene variants when performing metagenomic shotgun sequencing. Additionally, due to the longer nanopore sequencing reads, we can gain insights into specific AMRs and genomic context, as in the case study. In both the case study and the peri-rectal shotgun sequencing experiment, we were able to detect carbapenem resistant genes within 15 minutes after the start of sequencing, which can be useful information in regards to providing rapid, more targeted treatment.

We found a single nanopore run provided sufficient coverage to call consensus and fully assemble genomes when performing whole genome sequencing of individual isolates of *K. pneumoniae*. In addition to assembly, we were able to identify AMR genes, and examine the context of these genes in terms of location of the genome, flanking elements etc. We found the genes, number of plasmids and MLST typing from the Nanopore assemblies after error correction to be comparable to the Illumina assemblies; however, we were unable to correctly identify the type of specific genes pre-polishing. Through sequencing influenza, we have also characterized sample SNPs relative to a reference genome and to each other on both Illumina and Nanopore. When looking at influenza data, variant calling remains less accurate in the

Nanopore platform compared to the Miseq; nevertheless, we have found marked improvements in noise reduction after recent upgrades to R9.4 platform.

Future Directions of Nanopore Sequencing

Although nanopore sequencing has continually improved over the past two years, there remain areas available for improvement. While the Nanopore may have a relatively low instrumental overhead cost, the cost per sample is still higher than other sequencing platforms. This was particularly notable with previous iterations of the MinION chemistry, where the amount of sequencing data retrieved often varied and multiplexing samples would not have provided enough feasible data/coverage for purposes such as genome assembly. Current yields have increased dramatically - users have been achieving about 10Gb of data per flowcell, and newer high throughput systems Oxford Nanopore is developing, such as the GridION, could generate 100GB of data (Staff, n.d.). These yields will help drive down costs as MinION sequencing technology improves.

Data storage remains another challenge for the nanopore platform. Unlike traditional sequencing data file formats, such as fasta and fastq, the fast5 file produced for each nanopore sequencing read stores an entire hierarchical file structure. It contains bulky metadata files, containing information such as sequencing start times and electrical current information as the read was going through the sequencer. Some of this information can be useful later on, but it greatly expands file size, which is problematic for data storage, transfer, and pipeline runtimes. As Nanopore technology improves and more yield is generated, finding space to store and analyze all these data can become quite challenging and accrue more costs in the form of time, compute power, or storage. This detracts from the accessibility and flexibility benefits of the sequencer as datasets are shifting towards potentially becoming too large for a laptop computer

or even a powerful server. Because nanopore data can be analyzed in real time, eventually we can reach the point where this raw data can be deleted as soon as it is processed. This method is still in development but is promising (Stephens et al., 2015).

Although the consistency of flowcell quality and yield has increased dramatically with the progression of sequencing chemistries, nanopore sequencing can still be inconsistent in performance. Each flowcell contains 2048 pores, with typical numbers of active pores around 1100-1400. Although this is much improved over R7 chemistries, where flowcells ranged from 600-1000 pores, there is still room for improvement. Part of the inconsistency occurs due to the biological nature of the pores. As they are isolated from E. coli, these pore molecules are not necessarily robustly stable (Haque, Li, Wu, Liang, & Guo, 2013). One such way to get around this limitation is to create synthetic pores, providing simplified flowcell fabrication and nanopore consistency.. Work has already been done on creating solid-state pores, which allow for very exact production, and hybrid nanopores, which consist of a layer of alpha-hemolysin with the solid state nanopores, allowing for the advantages of both solid and biological nanopores (Hall et al., 2010; Haque et al., 2013).

Basecalling accuracy has also improved dramatically in recent years, and is actively being developed to further increase accuracy. Previous models used to translate electrical signal to nucleotide sequence were Hidden Markov Models, but more recent versions of Metrichor utilize a Recurrent Neural Network Model, which is more computationally intensive but also far more accurate. Currently, Oxford offers several basecallers that could be used to call data, including Albacore, the base code for Metrichor, and Nanonet, a developmental RNN. More validation needs to be performed on these programs in order to determine which one is the best method and most accurate program to use.

Finally, the need for error correctors that can correct basecalled reads and assemblies is also necessary. Currently, most error correction is performed with short read Illumina data and hybrid data software, such as through pilon. One error corrector, nanopolish, actually utilizes the raw electrical signal from the sequencing data, but many error correctors do not. Using the raw electrical data could have a lot of potential for other applications, especially as the raw data alone is nearly 4-5 times smaller in size than basecalled fast5 files, making files much easier to deal with and analysis a bit more rapid.

Other Considerations

Although Nanopore sequencing provides a platform and library preparations capable of going from extracted DNA to genomic sequence data in less than an hour, deficiencies still remain in DNA extraction methodologies. Many extraction protocols require mechanical lysis through bead beating. This method breaks apart not only the cell nucleus, but also severely fragments DNA as well. This could be detrimental to projects requiring long length reads, such as genome assembly and characterization. Methods that do not rely on beads can often take longer, and require more input DNA as well. For example, performing a Blue Pippin size selection to enrich for long reads can add another additional 3-4 hours to the protocol, negating some time saved from rapid sequencing. Additionally, there is the problem of preparing low-input samples for sequencing. Typically, protocols that involve low amounts of DNA require amplification to increase input, which can potentially introduce more biases into the sample.

In addition, pathogens do not make up a large proportion of organisms found in clinical samples. As shown with the chapter on vancomycin resistance, VRE organisms typically accounted for a small portion of the amount of reads sequenced, and antibiotic resistance genes

were detected in some samples in less than 1/1000 of the total reads. Determining whether these sequences are false positives can be challenging. Potentially, samples that are identified as positive for VRE by shotgun sequencing but have very few number of sequences that align back to a particular resistance could be subjected to further testing, such as targeted amplification specifically for the gene. More studies will need to be performed in order to determine a baseline threshold between calling a sample to be positive for specific AMR genes.

Overall, nanopore sequencing has been shown to be effective at detecting pathogenic organisms, and is making great progress towards detecting antimicrobial resistance genes. Its low capital cost and portability make it very attractive towards using it both in a clinical setting and out in the field at locations of outbreak. As improvements continue to be made in both yield and accuracy, we expect nanopore sequencing to become more useful as a diagnostic aid.

References:

Hall, A. R., Scott, A., Rotem, D., Mehta, K. K., Bayley, H., & Dekker, C. (2010). Hybrid pore formation by directed insertion of [alpha]-haemolysin into solid-state nanopores. *Nature Nanotechnology*, *5*(12), 874–877.

Haque, F., Li, J., Wu, H.-C., Liang, X.-J., & Guo, P. (2013). Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA. *Nano Today*, *8*(1), 56–74.

Staff, BioIT World. March 3, 2017. Oxford Nanopore Announces New Sequencing System, Certification Program. Retrieved May 3, 2017, from http://www.bio-itworld.com/2017/3/15/oxford-nanopore-announces-new-sequencing-system-certification-program.aspx

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., … Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biology*, *13*(7), e1002195.

Biographical Sketch

Stephanie Hao was born in 1993 in Seattle, Washington. She received her Bachelor's of Science in Biomedical Engineering in 2015 from Johns Hopkins, and her Master's of Science and Engineering in 2017 from Johns Hopkins.